

Hilbert Curve Projection Distance for Distribution Comparison

Tao Li, Cheng Meng, Hongteng Xu, *Member, IEEE*, Jun Yu

Abstract—Distribution comparison plays a central role in many machine learning tasks like data classification and generative modeling. In this study, we propose a novel metric, called *Hilbert curve projection (HCP) distance*, to measure the distance between two probability distributions with low complexity. In particular, we first project two high-dimensional probability distributions using Hilbert curve to obtain a coupling between them, and then calculate the transport distance between these two distributions in the original space, according to the coupling. We show that HCP distance is a proper metric and is well-defined for probability measures with bounded supports. Furthermore, we demonstrate that the modified empirical HCP distance with the L_p cost in the d -dimensional space converges to its population counterpart at a rate of no more than $O(n^{-1/2 \max\{d,p\}})$. To suppress the curse-of-dimensionality, we also develop two variants of the HCP distance using (learnable) subspace projections. Experiments on both synthetic and real-world data show that our HCP distance works as an effective surrogate of the Wasserstein distance with low complexity and overcomes the drawbacks of the sliced Wasserstein distance. The code of this work is at <https://github.com/sherlockLitao/HCP>.

Index Terms—Distribution comparison, optimal transport, Hilbert curve, Wasserstein distance, projection robust Wasserstein distance.

1 INTRODUCTION

MEASURING the distance between two probability distributions is significant for many machine learning tasks, e.g., data classification [1; 2; 3], generative modeling [4; 5], among others. Among the commonly-used distance measures for probability distributions, classic f -divergence based metrics, e.g., the Kullback-Leibler (KL) divergence and the total variation (TV) distance, do not work well when the probability distributions have disjoint supports [6], while the kernel-based methods like the maximum mean discrepancy (MMD) [7] require sophisticated kernel selection. Recently, the Wasserstein distance [8] has attracted wide attention in the machine learning community because of its advantages on overcoming these limitations, and it has shown great potential in many challenging learning problems [6; 9].

Given the samples of the two distributions, the computation of Wasserstein distance corresponds to solving either differential equations [10; 11] or linear programming problems [12; 13]. To alleviate the computational burden, the Sinkhorn distance [14] imposes an entropic regularizer on the Wasserstein distance and leverages the Sinkhorn-scaling algorithm accordingly. The work in [6] considers the Kantorovich duality of Wasserstein distance and converts the problem to a “max-min” game. Besides these two approximation methods, more surrogates of the Wasserstein distance

have been proposed in recent years, e.g., the sliced Wasserstein (SW) distance [15], the generalized sliced Wasserstein (GSW) distance [16], the tree-structured Wasserstein (TSW) distance [17], and so on. Despite the computational efficiency, these surrogates may fail to provide effective approximations for the Wasserstein distance. Take the two Gaussian mixture distributions in Fig. 1(a) as an example. We keep the source distribution (in purple) unchanged while shifting the central Gaussian component of the target distribution (in orange) vertically with an offset $\alpha \in [0, 1]$. For the various distances defined between the two distributions, Fig. 1(b) shows their changes with respect to α . Existing methods often lead to coarse approximations of the Wasserstein distance, whose tendencies w.r.t. α can even be opposite to the Wasserstein distance. This phenomenon indicates that replacing the Wasserstein distance with these surrogates may lead to sub-optimal, even undesired, results in some learning tasks.

In this study, we propose a novel metric for distribution comparison, called Hilbert curve projection (HCP) distance. In principle, our HCP distance first projects two probability distributions along the Hilbert curve [18] of the sample space and then calculates the coupling based on the projected distributions. Such a Hilbert curve projection works better than linear projections on preserving the structure of the data distribution since the Hilbert curve enjoys the locality-preserving property, i.e., the locality between data points in the high-dimensional space being preserved in the projected one-dimensional space [19; 20]. Our HCP distance provides a new surrogate of the Wasserstein distance, with both efficiency and effectiveness — it performs similarly as the Wasserstein distance does and spends less time than other methods, as shown in Fig. 1(b) and 1(c), respectively.

We provide in-depth analysis of the HCP distance, demonstrating that it is a well-defined metric for probability measures with bounded supports. Given n samples in d -dimensional space, the computational complexity for

- Tao Li is with the Institute of Statistics and Big Data, Renmin University of China.
E-mail: 2019000153@ruc.edu.cn
- Cheng Meng is with the Center for Applied Statistics, Institute of Statistics and Big Data, Renmin University of China.
E-mail: chengmeng@ruc.edu.cn
- Hongteng Xu is the corresponding author. He is with the Gaoling School of Artificial Intelligence, Renmin University of China and Beijing Key Laboratory of Big Data Management and Analysis Methods.
E-mail: hongtengxu@ruc.edu.cn
- Jun Yu is with the School of Mathematics and Statistics, Beijing Institute of Technology.
E-mail: yujunbeta@bit.edu.cn

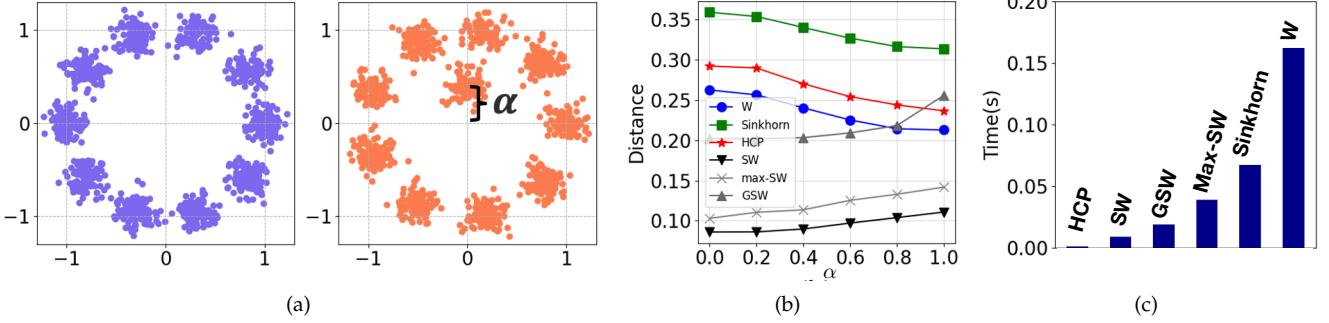


Fig. 1. (a) The samples of source and target distributions. (b) Illustrations of various distances with the increase of α . (c) Comparisons for various distances on their runtime. The proposed HCP distance provides an effective and efficient surrogate of the Wasserstein distance, which performs similarly and has low computational complexity.

calculating empirical HCP distance is approximately linear to n . In addition, the modified empirical HCP distance with the L_p cost converges to its population counterpart at a rate of no more than $O(n^{-1/2 \max\{d,p\}})$. Furthermore, to mitigate the curse-of-dimensionality, we develop two variants of the HCP distance using (learnable) subspace projections. We test the HCP distance and its variants on various machine learning tasks, including data classification and generative modeling, and compare them with state-of-the-art methods. Empirical results support the superior performance of the proposed metrics in both synthetic and real-data settings.

2 RELATED WORK AND PRELIMINARIES

2.1 Wasserstein distance and sliced Wasserstein distance

Let $\mathcal{P}_p(\mathbb{R}^d)$ be the set of Borel probability measures in \mathbb{R}^d with finite p -th moment. Consider two probability measures $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ with corresponding probability density functions f_μ, f_ν . The p -Wasserstein distance [8] between μ and ν is defined as

$$\begin{aligned} W_p(\mu, \nu) &= \left(\inf_{\gamma \in \Gamma(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|_p^p d\gamma(x, y) \right)^{1/p} \\ &\xrightarrow{\text{1D } \mu, \nu} \left(\int_0^1 \|F_\mu^{-1}(z) - F_\nu^{-1}(z)\|_p^p dz \right)^{1/p}, \end{aligned} \quad (1)$$

where $\|\cdot\|_p$ is the L_p norm and $\Gamma(\mu, \nu)$ is the set of all couplings (or called transportation plans): $\Gamma(\mu, \nu) = \{\gamma \in \mathcal{P}_p(\mathbb{R}^d \times \mathbb{R}^d) \text{ s.t. } \forall \text{ Borel set } A, B \subset \mathbb{R}^d, \gamma(A \times \mathbb{R}^d) = \mu(A), \gamma(\mathbb{R}^d \times B) = \nu(B)\}$.

Though it is difficult to calculate Wasserstein distance in general, according to (1), for one-dimensional probability measures μ and ν , the Wasserstein distance has a closed-form, where $F_\mu(x) = \mu((-\infty, x]) = \int_{-\infty}^x f_\mu(x) dx$ is the cumulative distribution function (CDF) for f_μ , and similarly, F_ν is the CDF for f_ν . This fact motivates the design of the sliced Wasserstein (SW) distance [15] (and its variants [21; 22]), which projects d -dimensional probability measures to 1D space and computes the 1D Wasserstein distance accordingly. Let $\mathbb{S}_{d,q} = \{\mathbf{E} \in \mathbb{R}^{d \times q} : \mathbf{E}^\top \mathbf{E} = \mathbf{I}_q\}$ ($q < d$) be the set of orthogonal matrices and $P_{\mathbf{E}}(\mathbf{x}) = \mathbf{E}^\top \mathbf{x}$ be the linear transformation for $\mathbf{x} \in \mathbb{R}^d$. Denote $P_{\mathbf{E}\#}\mu$ as the pushforward of μ by $P_{\mathbf{E}}$, which corresponds to the

distribution of the projected samples. For all $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, the p -sliced Wasserstein distance between them is given by

$$SW_p(\mu, \nu) = \left(\int_{\mathbf{E} \in \mathbb{S}_{d,1}} W_p^p(P_{\mathbf{E}\#}\mu, P_{\mathbf{E}\#}\nu) d\sigma(\mathbf{E}) \right)^{1/p}, \quad (2)$$

where σ is the uniform distribution on $\mathbb{S}_{d,1}$. However, as aforementioned, the SW distance often fails to approximate the Wasserstein distance because its linear projections break the structure of the original distributions. Additionally, the random projections introduce unnecessary randomness when computing the distance.

2.2 Other optimal transport distances

Based on the Wasserstein distance and the SW distance mentioned above, many optimal transport-based distances have been proposed in recent years, which can be roughly categorized into two classes. The first class considers approximating the Wasserstein distance by alternative optimization methods. Typically, the Sinkhorn distance in [14] imposes an entropic regularizer on the Wasserstein distance. Following this framework, many variants have been proposed to accelerate the computation [23; 24; 25; 26; 27; 28]. Besides the Sinkhorn-based algorithm, other methods, such as primal-dual method [29; 30], stochastic gradient descent [31], proximal point method [32], Bregman alternating direction method of multipliers (Bregman ADMM) [33; 34; 35], and so on, have drawn great attention. However, the computational cost of these methods is at least $O(n^2)$, which may not be applicable to large-scale data.

The second class follows the strategy of the SW distance, finding surrogates of the Wasserstein distance by various projection methods. To improve the efficiency of the SW distance, Max-sliced Wasserstein (Max-SW) [21], distributional sliced Wasserstein [36] and orthogonal sliced Wasserstein [37] have been proposed. Recently, the projection robust and integral projection robust Wasserstein distance consider projecting on the subspaces of higher dimensions [38; 39; 40]. Beyond using linear projections, generalized sliced Wasserstein (GSW) [16], convolutional sliced Wasserstein [41], and amortized sliced Wasserstein [42] have been proposed, which capture the complicated structure of data distributions by nonlinear projections. In addition to above methods, the tree sliced Wasserstein (TSW) [17] generates random tree metrics for data points and then computes Wasserstein distance on tree

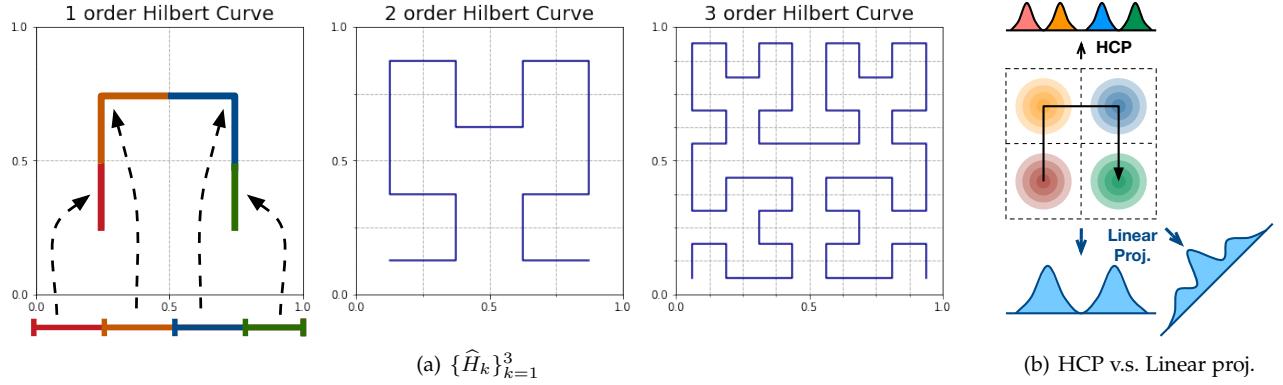


Fig. 2. (a) The k -order Hilbert curve, with $k = 1, 2, 3$, in 2D space. (b) The comparison between Hilbert curve projection (HCP) and linear projections.

metrics. TSW computes distance on given tree metrics and thus, it is more suitable for classification task compared with generative model. Note that, these projection-based distances may fail to provide effective surrogates for the Wasserstein distance, as shown in Fig. 1, and searching for effective projections will bring additional computational cost.

2.3 Applications of optimal transport distances

Optimal transport distances have recently drawn great attention in various machine-learning tasks. Wasserstein distance and its variants serve as the loss functions for generative modeling, such as Wasserstein generative adversarial networks (WGANs) [6; 21; 43; 44] and Wasserstein autoencoders (WAEs) [9; 22; 45]. In classification tasks, optimal transport distances measure the discrepancy between set-level data [46], leading to discriminative models for various data, such as texts [1; 2], point clouds [3], and graphs [47]. Besides generative modeling and classification, optimal transport distances are also applied to other problems, such as data clustering [34; 48], dimension reduction [49; 50], and domain adaptation [51]. These optimal transport-based methods have shown the potential for various practical applications, e.g., graph matching and partitioning [52; 53], color transfer [54; 55], document analysis [56], and so on.

3 PROPOSED METHOD

3.1 Hilbert curve and its locality-preserving property

Our work is based on the well-known Hilbert curve [18]. Mathematically, for a d -dimensional ($d \geq 2$) unit hyper-cube, i.e., $[0, 1]^d$, the k -order Hilbert space-filling curve, denoted as \hat{H}_k , partitions $[0, 1]$ and $[0, 1]^d$ into $(2^k)^d$ intervals and blocks, respectively, and constructs a bijection between them. Taking the $\{\hat{H}_k\}_{k=1}^3$ in 2D space as examples, Fig. 2(a) illustrates how the intervals in $[0, 1]$ are constructed and mapped to the blocks in $[0, 1]^2$. The Hilbert curve is defined as the limit of a sequence of k -order Hilbert space-filling curves, i.e., $H(x) = \lim_{k \rightarrow \infty} \hat{H}_k(x)$ with $x \in [0, 1]$. It provides a well-defined surjection $H : [0, 1] \rightarrow [0, 1]^d$ and is able to cover the entire hyper-cube [18]. Note that, although the Hilbert curve H is not a bijection, most of the data points in $[0, 1]^d$ are still invertible — it is known that the set \mathcal{A} , which includes the points in $[0, 1]^d$ such that these points have more than one pre-image in $[0, 1]$, has measure zero [57]. Actually, for any

point in \mathcal{A} , there are finite pre-images in $[0, 1]$. Hence, there is a bijection between $[0, 1]^d$ and $\{\min\{H^{-1}(x)\} : x \in [0, 1]^d\}$. We denote $\mathcal{N} = [0, 1] \setminus \{\min\{H^{-1}(x)\} : x \in [0, 1]^d\}$ as the negligible set.

We are interested in the Hilbert curve because it enjoys the so-called **locality-preserving property** [57; 58]: For any $x, y \in [0, 1]$, one has

$$\|H(x) - H(y)\|_2 \leq 2\sqrt{d+3}|x - y|^{1/d}.$$

Such an inequality indicates the advantage of the Hilbert curve over linear projections. In particular, if two points are far from each other in a high-dimensional space, their pre-images with respect to the Hilbert curve will also be far from each other. Fig. 2(b) further illustrates this property through a toy example. Specifically, for a 2D distribution with four modals, while linear projections tend to wrongly merge some modals, the projection along the Hilbert curve can distinguish the modals successfully. This property motivates us to propose the Hilbert curve projection distance shown below.

3.2 Hilbert curve projection distance

Hilbert curve for probability measure: In this study, we focus on probability measures with bounded supports. This condition has been widely used in the optimal transport literature to simplify the theoretical analysis [59; 60; 61; 62]. Let $\mathcal{P}_\infty(\mathbb{R}^d)$ be the set of Borel probability measures in \mathbb{R}^d with bounded supports. Denote the support of a probability measure $\mu \in \mathcal{P}_\infty(\mathbb{R}^d)$ as Ω_μ . Let $\tilde{\Omega}_\mu = \prod_{i=1}^d [a_i, b_i]$ be the smallest hyper-rectangle covering Ω_μ . For each μ , we can define a Hilbert curve $H_\mu : [0, 1] \rightarrow \tilde{\Omega}_\mu$ as $H_\mu(t) = (b - a) \odot H(t) + a$ where \odot is the Hadamard product and a, b are vectors with i -th dimension being a_i, b_i respectively.¹

Denote $\mathcal{K} = \{\frac{m_1}{2^{m_2}} : m_1, m_2 \in \mathbb{N}, m_1 \leq 2^{m_2}\}$ as a dense set in $[0, 1]$. According to [57; 58], $H_\mu([0, t])$ is a Borel measurable set for any $t \in \mathcal{K}$ and $H_\mu([0, t])$ is a Lebesgue measurable set for any $t \in [0, 1]$. This motivates us to define

1. In the case when $a_i = b_i$, one can utilize the following two strategies without affecting the theoretical properties. The first strategy is to let $b_i = a_i + 1$. This may cause redundant computational costs in this dimension. The second strategy is removing this dimension, performing Hilbert curve in the \mathbb{R}^{d-1} and complementing this dimension for the final Hilbert curve.

a cumulative distribution function along the Hilbert curve (denoted as $g_\mu : [0, 1] \rightarrow [0, 1]$) and the corresponding inverse cumulative distribution function (g_μ^{-1}), respectively:

$$\begin{aligned} g_\mu(t) &= \inf_{s \in \mathcal{K}, s \geq t} \mu(H_\mu([0, s])), \\ g_\mu^{-1}(t) &= \inf_{s \in [0, 1], g_\mu(s) > t} s. \end{aligned} \quad (3)$$

Accordingly, the formal definition of our Hilbert curve projection distance is as follows.

Definition 1 (Hilbert Curve Projection Distance). Let $\mathcal{P}_\infty(\mathbb{R}^d)$ be the set of Borel probability measures in \mathbb{R}^d with bounded supports. Denote the supports of two probability measures $\mu, \nu \in \mathcal{P}_\infty(\mathbb{R}^d)$ as Ω_μ and Ω_ν , respectively. Denote $\mathcal{K} = \{\frac{m_1}{2^{m_2}} : m_1, m_2 \in \mathbb{N}, m_1 \leq 2^{m_2}\}$ as a dense set in $[0, 1]$. Let $H_\mu : [0, 1] \rightarrow \tilde{\Omega}_\mu$, where $\tilde{\Omega}_\mu$ is the smallest hyper-rectangle that covers Ω_μ , $g_\mu(t) = \inf_{s \in \mathcal{K}, s \geq t} \mu(H_\mu([0, s]))$, and $g_\mu^{-1}(t) = \inf_{s \in [0, 1], g_\mu(s) > t} s$ (with H_ν, g_ν and g_ν^{-1} defined in the same way). For $p \in \mathbb{Z}_+$, the p -order Hilbert curve projection distance is defined as

$$HCP_p(\mu, \nu) = \left(\int_0^1 \|H_\mu(g_\mu^{-1}(t)) - H_\nu(g_\nu^{-1}(t))\|_p^p dt \right)^{\frac{1}{p}}. \quad (4)$$

Remark 1. The assumption for bounded support is commonly used in optimal transport literature [59; 60; 61; 62] and is essential for technical proof. For unbounded cases, one possible remedy is to use a bounded measurable bijective mapping f , such as element-wise $\tan^{-1}(\cdot)$, to transform the original measures μ and ν . We then could get the transport plan between $f_\# \mu$ and $f_\# \nu$ based on the Hilbert curve projections where $f_\# \mu$ is the pushforward of μ by f , and compute the distance in the original unbounded space.

According to the definition, the principle of our HCP distance is projecting high-dimensional distributions along their Hilbert curves to obtain an efficient and effective coupling between them, and then calculating the corresponding HCP distance between two distributions in the original space according to the coupling. The following theoretical results show that our HCP distance is a proper metric, and it is an upper bound of the p -Wasserstein distance.

Theorem 1. HCP_p is a well-defined metric in $\mathcal{P}_\infty(\mathbb{R}^d)$, and $W_p(\mu, \nu) \leq HCP_p(\mu, \nu), \forall \mu, \nu \in \mathcal{P}_\infty(\mathbb{R}^d)$.

Given two random variables, i.e., $Z_1 \sim \mu$ and $Z_2 \sim \nu$, we denote $HCP(\mu, \nu)$ as $HCP(Z_1, Z_2)$. Clearly, HCP distance has the following properties which are also valid for Wasserstein distance [63].

- 1) For any $z \in \mathbb{R}^d$, $HCP_p(Z_1 + z, Z_1) = \|z\|_p$.
- 2) For any $a \in \mathbb{R}$, $HCP_p(aZ_1, aZ_2) = |a|HCP_p(Z_1, Z_2)$.
- 3) For any $z \in \mathbb{R}^d$, $HCP_p(Z_1 + z, Z_2 + z) = HCP_p(Z_1, Z_2)$.
- 4) For any $z \in \mathbb{R}^d$, $HCP_2^2(Z_1 + z, Z_2) = HCP_2^2(Z_1, Z_2) + \|z + \mathbb{E}Z_1 - \mathbb{E}Z_2\|_2^2 - \|\mathbb{E}Z_1 - \mathbb{E}Z_2\|_2^2$.

Here, “ $Z_1 + z$ ” means impose a translation z on the random variable Z_1 , and “ aZ_1 ” means scaling the random variable Z_1 .

3.3 Topological properties of the HCP distance

As shown in Theorem 1, HCP distance induces a stronger topology compared to Wasserstein distance because $W_p(\mu, \nu) \leq HCP_p(\mu, \nu)$. This means that the sequence

of probability measures, i.e., $\{\mu_n\}$, always converges in Wasserstein distance when $n \rightarrow \infty$ if it converges in HCP distance, i.e., $HCP(\mu_n, \mu) \rightarrow 0 \Rightarrow W(\mu_n, \mu) \rightarrow 0$.

Additionally, we compare our HCP distance with the total variation (TV) distance on their induced topology and propose the following Theorem:

Theorem 2. Let $\tilde{\Omega}_\mu$ be the smallest hyper-rectangle that covers the support of the probability measure $\mu \in \mathcal{P}_\infty(\mathbb{R}^d)$. When $\{\mu_n\}$ converges to μ in the total variation distance and $\Omega_{\mu_n} = \tilde{\Omega}_\mu$ for all n 's, we have $TV(\mu_n, \mu) \rightarrow 0 \Rightarrow HCP(\mu_n, \mu) \rightarrow 0$.

Note that, our HCP distance is not equivalent to the Wasserstein distance or the TV distance because

$$\begin{aligned} W(\mu_n, \mu) \rightarrow 0 &\not\Rightarrow HCP(\mu_n, \mu) \rightarrow 0, \\ HCP(\mu_n, \mu) \rightarrow 0 &\not\Rightarrow TV(\mu_n, \mu) \rightarrow 0. \end{aligned}$$

The following two examples verify the above claims, respectively.

Example 1. Consider two probability distribution $\mu_\theta = \frac{1}{4}(\delta_{(0,0)} + \delta_{(1,1)} + \delta_{(\frac{1}{2}-\theta, \frac{1}{4})} + \delta_{(\frac{1}{2}-\theta, \frac{3}{4})})$ and $\nu_\theta = \frac{1}{4}(\delta_{(0,0)} + \delta_{(1,1)} + \delta_{(\frac{1}{2}+\theta, \frac{1}{4})} + \delta_{(\frac{1}{2}+\theta, \frac{3}{4})})$ where δ is the Dirac measure. Then, when $0 < \theta < 0.5$, we have $W_2(\mu_\theta, \nu_\theta) = |\sqrt{2}\theta|$. However, when $\theta \neq 0$, $HCP_2(\mu_\theta, \nu_\theta) = \sqrt{2\theta^2 + 1/8}$.

Example 2. Let $Z \sim \text{Unif}[0, 1]$ be samples of the uniform distribution on the unit interval. Let μ_0 be the probability distribution of $(0, Z) \in \mathbb{R}^2$. Let μ_θ be the family of probability distributions parametrized with θ corresponding to $(\theta, Z) \in \mathbb{R}^2$. Then $HCP_p(\mu_0, \mu_\theta) = W_p(\mu_0, \mu_\theta) = |\theta|$. However, when $\theta \neq 0$, $TV(\mu_0, \mu_\theta) = 1$.

In summary, we can find that HCP metricizes a topology stronger than the weak topology induced by the Wasserstein distance. Additionally, as shown in Example 2 (which is also used in [6]), our HCP distance can perform as well as the Wasserstein distance does when comparing the probability measures with disjoint supports.

3.4 Numerical implementation

Let Δ^n be the n -Simplex. Given the samples of two probability measures, i.e., $X = \{x_i\}_{i=1}^n \sim \mu$ and $Y = \{y_j\}_{j=1}^m \sim \nu$, whose empirical distributions are $\mathbf{a} \in \Delta^{n-1}$ and $\mathbf{b} \in \Delta^{m-1}$, respectively, we use a k -order Hilbert curve to calculate the empirical HCP distance between the two sample sets. Let $\tilde{\Omega}_X$ and $\tilde{\Omega}_Y$ be the smallest hyper-rectangles that cover these two sample sets, respectively. We define two k -order Hilbert curves, i.e., $\hat{H}_k^X : [0, 1] \rightarrow \tilde{\Omega}_X$ and $\hat{H}_k^Y : [0, 1] \rightarrow \tilde{\Omega}_Y$. Here, \hat{H}_k^X partitions both $[0, 1]$ and $\tilde{\Omega}_X$ into 2^{kd} blocks, denoted by $\{c'_{j,X}\}_{j=1}^{2^{dk}}$ and $\{c_{j,X}\}_{j=1}^{2^{dk}}$, respectively, and construct a bijection between these blocks. For any data point $x \in \tilde{\Omega}_X$, we assign x to its corresponding block $c_{j,X}$ in $\tilde{\Omega}_X$, $j \in \{1, \dots, 2^{kd}\}$, then map x to the center of the block $c'_{j,X} = (\hat{H}_k^X)^{-1}(c_{j,X})$. Therefore, all the samples belonging to the same block are mapped to the same point in $[0, 1]$. Based on \hat{H}_k^Y , we map $\{y_j\}_{j=1}^m$ to $[0, 1]$ in the same way. The mapped points along with their probability densities are then used to calculate the optimal coupling matrix $\mathbf{P} \in \mathbb{R}^{n \times m}$ using the closed-form formulation of the 1D optimal transport problem. In particular, we first sort the mapped points, then calculate \mathbf{P} using the North-West corner

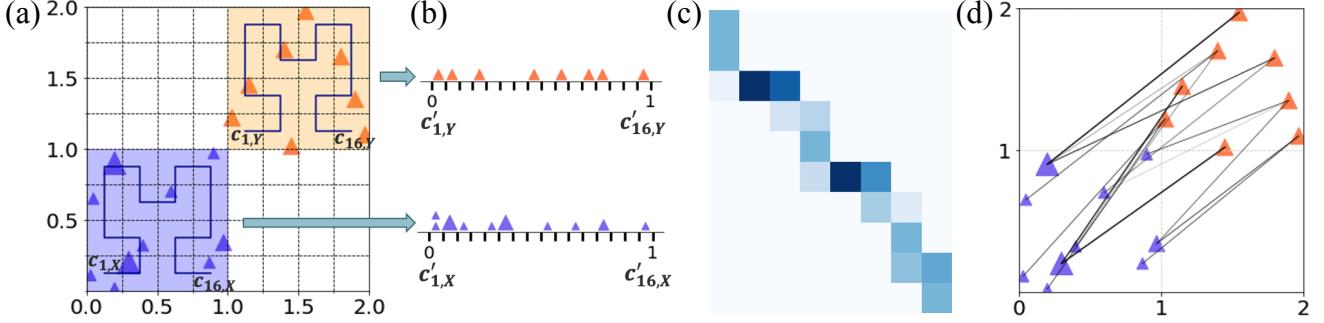


Fig. 3. An illustration of Algorithm 1 when $d = k = 2$. (a) The source (purple) and target (orange) data points, with corresponding hyper-rectangles and k -order Hilbert curves. (b) The projected points along the Hilbert curves. (c) The coupling matrix calculated by the projected points. (d) The HCP distance calculates the distance between the original samples based on the coupling matrix.

Algorithm 1 Computation of HCP distance

- 1: **Input:** $(\{x_i\}_{i=1}^n, \mathbf{a}), (\{y_j\}_{j=1}^m, \mathbf{b}), k$
- 2: Map $\{x_i\}_{i=1}^n$ to $\{x'_i\}_{i=1}^n, \{y_j\}_{j=1}^m$ to $\{y'_j\}_{j=1}^m$, through $(\hat{H}_k^X)^{-1}$ and $(\hat{H}_k^Y)^{-1}$ $O((n+m)dk)$
- 3: Calculate the optimal transport plan \mathbf{P} between $(\{x'_i\}_{i=1}^n, \mathbf{a})$ and $(\{y'_j\}_{j=1}^m, \mathbf{b})$ using sorting and the North-West corner rule. Let $\mathcal{S} := \{(i, j) | P_{ij} \neq 0\}$ $O(n \log(n) + m \log(m))$
- 4: **Output:** $\mathbf{P}, \text{HCP}_p = (\sum_{(i,j) \in \mathcal{S}} \|x_i - y_j\|_p^p P_{ij})^{1/p}$

rule with $O(n + m)$ operations [64]. Note that there are at most $m + n$ nonzero elements in \mathbf{P} . Let $\mathcal{S} := \{(i, j) | P_{ij} \neq 0\}$ be the index set. Finally, the empirical HCP distance can then be calculated by $(\sum_{(i,j) \in \mathcal{S}} \|x_i - y_j\|_p^p P_{ij})^{1/p}$.

The above pipeline is illustrated in Fig. 3 and summarized in Algorithm 1, respectively. As suggested by [18], we select k that of the order $O(\log(n))$ in practice. Empirical results in the following experimental section show that the performance of Algorithm 1 is not sensitive to k .

Essentially, the empirical HCP distance is to compute the distance between two Hilbert rank-based sorted samples.² Note that, there are two main routines for Hilbert sort. The first gets Hilbert indices by projecting points in high dimension to the Hilbert curve and then sorts these indices based on the Hilbert rank [18; 58; 65; 66; 67]. The second idea is recursively sorting points without using Hilbert indices, e.g., the work in [18; 68; 69] and the C++ library CGAL [70]. Though we take the first routine here, codes based on these two algorithms are both provided.

Computational cost. The complexity of computing the k -order Hilbert index for n points in d -dimensional space is $O(ndk)$ [68; 69; 71]. As shown in Algorithm 1, solving the optimal transport problem in Step 3 requires $O(n \log n + m \log m)$ time. When $m = O(n)$ and $k = O(\log(n))$, the overall computational complexity of HCP distance is at the order of $O(n \log(n)d)$.

Comparison with existing methods. The proposed HCP distance enjoys several critical advantages over the Wasserstein and SW distance.

- Firstly, HCP can provide a decent transport plan between the input probability measures as a byproduct

2. The Hilbert rank is defined as follows: We say x_1 ranks in front of x_2 , that is to say, $\min\{H^{-1}(x_1)\} < \min\{H^{-1}(x_2)\}$.

while SW could not. The key reason is that Hilbert curve is invertible almost everywhere. Linear projections in SW and nonlinear projections in GSW do not satisfy this property. Such a coupling matrix is essential for effective generative modeling, as will be seen in Section 5.

- Secondly, we compute the distance in the original space rather than in the projected one-dimensional space. Hilbert curve only plays a role in achieving a transport plan. We don't apply any transformation on data points when computing HCP distance. However, SW involves transforming data points by linear projections and then computing Wasserstein distance using these transformed data points. Fig. 1 provides an intuitive example to show the difference between these two strategies. The reason why SW and its variants lead to an opposite trend compared with the Wasserstein distance is that SW computes Wasserstein distance using linear transformed data points, and such linear transformation may break the structure of the original distributions. We refer to the Experiment Section for a more intuitive discussion.
- Last but not least, HCP computes faster than SW distance in practice. This is because calculating SW distance requires projection and sorting multiple times, while calculating HCP distance requires only once. Additionally, beyond the Hilbert curve-based discrepancy in [72], our HCP distance can deal with the samples with different sizes and weights with theoretical guarantees.

In summary, compared to Wasserstein distance, HCP has an approximately linear computational complexity, and thus is applicable to large-scale datasets. Compared to SW distances, HCP distance performs more similarly to the Wasserstein distance.

3.5 Statistical convergence of empirical HCP distance

Let $\{x_i\}_{i=1}^n \sim \mu$, whose empirical measure is defined by $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$. Directly studying the statistical convergence of $\text{HCP}_p(\mu, \mu_n)$ is challenging because of the randomness of the bounded supports — the smallest hyper-rectangle covering the support of the probability measure μ_n , i.e., Ω_{μ_n} can be various w.r.t. sample size n , which leads to different Hilbert curves, and accordingly, we could not easily analyze the convergence rate without any other strict conditions on the support's boundary.

To eliminate the influence of the randomness, we consider an indirect strategy, studying a modified empirical Hilbert

curve projection distance instead. Specifically, following the definitions in (3), we first define the cumulative distribution function and its inverse for the empirical measure μ_n , whose Hilbert curve, however, is based on the original probability measure μ :

$$\begin{aligned}\hat{g}_{\mu_n}(t) &= \inf_{s \in \mathcal{K}, s \geq t} \mu_n(H_\mu([0, s])), \\ \hat{g}_{\mu_n}^{-1}(t) &= \inf_{s \in [0, 1], \hat{g}_{\mu_n}(s) > t} s.\end{aligned}\quad (5)$$

Accordingly, we define the modified empirical Hilbert curve projection distance as:

$$\overline{\text{HCP}}_p(\mu, \mu_n) = \left(\int_0^1 \|H_\mu(g_\mu^{-1}(t)) - H_\mu(\hat{g}_{\mu_n}^{-1}(t))\|_p^p dt \right)^{\frac{1}{p}}. \quad (6)$$

The only difference between $\text{HCP}_p(\mu, \mu_n)$ and $\overline{\text{HCP}}_p(\mu, \mu_n)$ is that the latter replaces the H_{μ_n} defined on $\tilde{\Omega}_{\mu_n}$ with the H_μ defined on $\tilde{\Omega}_\mu$. Note that, such a modified HCP distance is hard to implement in practice because both $\tilde{\Omega}_\mu$ and H_μ are unknown in general. However, compared to the original HCP distance, the modified HCP distance is much easier to analyze because the Hilbert curve H_μ it used is deterministic and irrelevant to the sample. We demonstrate that the modified empirical HCP distance converges to its population counterpart almost surely. The following theorem provides an upper bound for the convergence rate.

Theorem 3. Let $\{x_i\}_{i=1}^n$ be an i.i.d. sample that is generated from the probability measure $\mu \in \mathcal{P}_\infty(\mathbb{R}^d)$. The empirical measure is defined by $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$. Then, we have almost surely

$$\overline{\text{HCP}}_p(\mu, \mu_n) \rightarrow 0, \text{ and } \mathbb{E} \overline{\text{HCP}}_p(\mu, \mu_n) \lesssim O(n^{-\frac{1}{2 \max\{p, d\}}}).$$

Directly from Theorem 3, we can conclude the following theoretical results.

Corollary 3.1. Assume that probability measures $\mu, \nu \in \mathcal{P}_\infty(\mathbb{R}^d)$. Let $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^m$ be two i.i.d. samples, which are generated from probability measures μ and ν , respectively. Let $\{x_{(i)*}\}_{i=1}^n$ and $\{y_{(i)*}\}_{i=1}^m$ be the sorted samples along the Hilbert curves H_μ and H_ν , respectively. Then, we have almost surely

$$\overline{\text{HCP}}_p(\mu_n, \nu_n) = \left(\frac{1}{n} \sum_{i=1}^n \|x_{(i)*} - y_{(i)*}\|_p^p \right)^{\frac{1}{p}} \rightarrow \text{HCP}_p(\mu, \nu),$$

where μ_n and ν_n are the empirical version of μ and ν , respectively. Furthermore, we have

$$|\mathbb{E} \overline{\text{HCP}}_p(\mu_n, \nu_n) - \text{HCP}_p(\mu, \nu)| \lesssim O(n^{-\frac{1}{2 \max\{p, d\}}}).$$

Corollary 3.1 tells us the modified empirical Hilbert curve distance is to compute the distance between two Hilbert rank-based sorted samples. Moreover, when the samples are with different numbers, we have

Corollary 3.2. Assume that probability measures $\mu, \nu \in \mathcal{P}_\infty(\mathbb{R}^d)$. Let $\{x_i\}_{i=1}^n$ and $\{y_j\}_{j=1}^m$ be two i.i.d. samples, which are generated from probability measures μ and ν , respectively. Let $\{x_{(i)*}\}_{i=1}^n$ and $\{y_{(j)*}\}_{j=1}^m$ be the sorted samples along the Hilbert curves H_μ and H_ν , respectively. Then, we have

$$\overline{\text{HCP}}_p(\mu_n, \nu_m) = \left(\pi_{ij} \sum_{i=1}^n \sum_{j=1}^m \|x_{(i)*} - y_{(j)*}\|_p^p \right)^{\frac{1}{p}},$$

where μ_n, ν_m are the empirical version of μ, ν , respectively and, π_{ij} is the optimal transport plan between $\sum_{i=1}^n \delta_i/n$ and $\sum_{j=1}^m \delta_j/m$ with Euclidean distance cost. Furthermore, we have

$$|\mathbb{E} \overline{\text{HCP}}_p(\mu_n, \nu_m) - \text{HCP}_p(\mu, \nu)| \lesssim O(\min\{n, m\}^{-\frac{1}{2 \max\{p, d\}}}).$$

Additionally, from Theorem 3, we know that convergence rate of modified empirical HCP distance has an upper bound $O(n^{-1/2p} + n^{-1/2d})$, which is slightly slower than the convergence rate of Wasserstein distance (i.e., $O(n^{-1/2p} + n^{-1/d})$) provided by [63]). In particular, given a probability measure μ and its empirical version μ_n , we have

$$W_p(\mu, \mu_n) \leq \overline{\text{HCP}}_p(\mu, \mu_n).$$

Furthermore, the following corollary indicates that under some mild conditions, the modified HCP distance can have the same convergence rate as Wasserstein distance does.

Corollary 3.3. Assume that probability measure $\mu \in \mathcal{P}_\infty(\mathbb{R}^d)$. If there exist two Borel measurable sets $A, B \subset \mathbb{R}^d$ such that $\mu(A) > 0, \mu(B) > 0, \mu(A \cup B) = 1$ and $\text{dist}(A, B) = \inf_{x \in A, y \in B} \|x - y\|_2 > 0$, then when $p \geq d$, we have

$$\mathbb{E} \overline{\text{HCP}}_p(\mu, \mu_n) = O(n^{-\frac{1}{2p}}), \text{ and } \mathbb{E} W_p(\mu_n, \mu) = O(n^{-\frac{1}{2p}}),$$

where μ_n is the empirical version of μ .

The above theoretical results of the modified HCP distance provide us with important insights into the convergence of our HCP distance — with the increase of the sample size n , the difference between $\tilde{\Omega}_{\mu_n}$ and $\tilde{\Omega}_\mu$ may not be too large in probability. Accordingly, the convergence of our HCP distance should be similar to that of the modified HCP distance in probability as well. Under some special cases, we can easily analyze the convergence of our HCP distance. For example, for nondegenerate discrete measures with finite supports, $\mathbb{E} \overline{\text{HCP}}_p(\mu, \mu_n)$ is of the order $O(n^{-1/2p})$, independently of the dimension, which is the same as the Wasserstein distance [63].

Corollary 3.4. Assume that probability measure μ is a non-degenerate discrete probability measure with K supports $\{s_i\}_{i=1}^K$, that is, $\mu = \sum_{i=1}^K p_i \delta_{s_i}$ and $\mathbf{p} = \{p_i\}_{i=1}^K \in \Delta^{K-1}$. Then, we have

$$\mathbb{E} \overline{\text{HCP}}_p(\mu, \mu_n) = O(n^{-\frac{1}{2p}}),$$

where μ_n is the empirical version of μ .

3.6 Other space-filling curves

The proposed distance can be implemented based on other space-filling curves as well. For example, the Peano and Sierpinski space-filling curves also satisfy the Hölder inequality with exponent $1/d$. The Z-order space-filling curve, which is differentiable almost everywhere, also satisfies the Hölder inequality but with exponent $1/(d \log_2 3)$ [57; 58]. However, compared to the Hilbert curve, the Peano curve and Sierpinski curve are difficult to implement through algorithms. The convergence rate of the distance based on the Z-order curve is $O(n^{-\frac{1}{2 \max\{d \log_2 3, p\}}})$, which is slower than that based on the Hilbert curve. In sum, we mainly focus on the Hilbert curve in this study.

4 VARIANTS OF THE HILBERT CURVE PROJECTION DISTANCE

The theoretical results in the previous section indicate that analogous to the Wasserstein distance, our HCP distance may suffer from the curse-of-dimensionality as well. Motivated by the projection-robust Wasserstein distance [39; 40], we propose two variants of the HCP distance to alleviate this limitation.

4.1 Integral projection robust Hilbert curve projection distance

We first propose the integral projection robust Hilbert curve projection (IPRHCP) distance that combines the idea of HCP distance and random projections.

Definition 2. Suppose that probability measures $\mu, \nu \in \mathcal{P}_\infty(\mathbb{R}^d)$. The p -order q -dimensional integral projection robust Hilbert curve projection distance is defined as

$$\begin{aligned} & \text{IPRHCP}_{p,q}(\mu, \nu) \\ &= \left(\int_{\mathbf{E} \in \mathbb{S}_{d,q}} HCP_p^p(P_{\mathbf{E}\#}\mu, P_{\mathbf{E}\#}\nu) d\sigma(\mathbf{E}) \right)^{\frac{1}{p}}, \end{aligned} \quad (7)$$

where σ is the uniform distribution on $\mathbb{S}_{d,q}$.

Next, we demonstrate that IPRHCP distance is a valid distance metric and reveal the relations between IPRHCP distance and other metrics, including the p -order SW distance [15] and the p -order q -dimensional integral projection robust Wasserstein distance [40], denoted as $\text{IPRW}_{p,q}$.

Theorem 4. $\text{IPRHCP}_{p,q}$ is a well-defined metric in $\mathcal{P}_\infty(\mathbb{R}^d)$, and we have $\text{IPRW}_{p,q}(\mu, \nu) \leq \text{IPRHCP}_{p,q}(\mu, \nu)$, $\forall \mu, \nu \in \mathcal{P}_\infty(\mathbb{R}^d)$.

In practice, the expectation in (7) can be approximated using a Monte Carlo scheme: We first randomly and uniformly draw several matrices from the set of orthogonal matrices $\mathbb{S}_{d,q}$. We then project the distributions to subspace \mathbf{E} and compute the HCP distance between the projected samples. Finally, we replace the expectation on the right hand side of (7) with a finite-sample average.

Theorem 5. Given two probability measures $\mu, \nu \in \mathcal{P}_\infty(\mathbb{R}^d)$, we have $\text{SW}_p^p(\mu, \nu) \leq \alpha_{q,p} \text{IPRHCP}_{p,q}^p(\mu, \nu)$, where $\alpha_{q,p} = \int_{\mathbb{S}_{q,1}} \|\theta\|_p^p d\theta / q \leq 1$. As a special case, when $p = 2$, one has $\alpha_{q,2} = 1/q$ and $\text{SW}_2(\mu, \nu) \leq \text{IPRHCP}_{2,q}(\mu, \nu) / \sqrt{q}$.

Corollary 5.1. If we replace $\mathbb{S}_{d,q}$ in (7) with matrix set $\{\mathbf{E} \in \mathbb{R}^{d \times q} : \mathbf{E}^\top \mathbf{E} = \mathbf{J}_q\}$ where \mathbf{J}_q is a $q \times q$ all-ones matrix, we have $\text{IPRHCP}_{p,q}(\mu, \nu) = q^{1/p} \text{SW}_p(\mu, \nu)$, $\forall \mu, \nu \in \mathcal{P}_\infty(\mathbb{R}^d)$.

IPRHCP shares a similar sense to SW. As shown in Theorem 5 and Corollary 5.1, we provided some inequalities and equalities between IPRHCP and SW to illustrate their relationship. We provide the following theorem to show IPRHCP overcomes curse-of-dimensionality.

Theorem 6. Suppose that probability measures $\mu, \nu \in \mathcal{P}_\infty(\mathbb{R}^d)$. Let $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ be two i.i.d. samples, which are generated from probability measures μ and ν , respectively. Let $\{x_{\mathbf{E},(i)^*}\}_{i=1}^n$ and $\{y_{\mathbf{E},(i)^*}\}_{i=1}^n$ be the sorted samples of $\{\mathbf{E}^\top x_i\}_{i=1}^n$ and

$\{\mathbf{E}^\top y_i\}_{i=1}^n$ along the Hilbert curves $H_{P_{\mathbf{E}\#}\mu}$ and $H_{P_{\mathbf{E}\#}\nu}$, respectively. Based on the definition of $\overline{\text{HCP}}(\mu_n, \nu_n)$, we can define

$$\begin{aligned} & \overline{\text{IPRHCP}}_{p,q}(\mu_n, \nu_n) \\ &= \left(\int_{\mathbf{E} \in \mathbb{S}_{d,q}} \overline{\text{HCP}}_p^p(P_{\mathbf{E}\#}\mu_n, P_{\mathbf{E}\#}\nu_n) d\sigma(\mathbf{E}) \right)^{\frac{1}{p}} \\ &= \left(\int_{\mathbf{E} \in \mathbb{S}_{d,q}} \frac{1}{n} \sum_{i=1}^n \|x_{\mathbf{E},(i)^*} - y_{\mathbf{E},(i)^*}\|_p^p d\sigma(\mathbf{E}) \right)^{\frac{1}{p}}, \end{aligned}$$

where μ_n and ν_n are the empirical version of μ and ν , respectively. Then, we have

$$|\mathbb{E} \overline{\text{IPRHCP}}_{p,q}(\mu_n, \nu_n) - \text{IPRHCP}_{p,q}(\mu, \nu)| \lesssim O(n^{-\frac{1}{2\max\{p,q\}}}).$$

4.2 Projection robust Hilbert curve projection distance

The IPRHCP distance considers the integration of the HCP distances defined in all q -dimensional subspaces. When assuming the two distributions differ only on one low-dimensional subspace, as the projection robust Wasserstein (PRW) distance [39] does, we can avoid the integration and just consider the maximal possible HCP distance among all projections, which leads to the proposed projection robust Hilbert curve projection (PRHCP) distance.

Definition 3. Suppose that probability measures $\mu, \nu \in \mathcal{P}_\infty(\mathbb{R}^d)$. The p -order q -dimensional projection robust Hilbert curve projection distance is defined as

$$\text{PRHCP}_{p,q}(\mu, \nu) = \sup_{\mathbf{E} \in \mathbb{S}_{d,q}} \text{HCP}_p(P_{\mathbf{E}\#}\mu, P_{\mathbf{E}\#}\nu). \quad (8)$$

The PRHCP distance is also a valid distance.

Theorem 7. $\text{PRHCP}_{p,q}(\mu, \nu)$ is a well-defined metric in $\mathcal{P}_\infty(\mathbb{R}^d)$, and we have $\text{PRW}_{p,q}(\mu, \nu) \leq \text{PRHCP}_{p,q}(\mu, \nu)$, $\forall \mu, \nu \in \mathcal{P}_\infty(\mathbb{R}^d)$.

In practice, given the samples of the probability measures, i.e., the sample matrices $\mathbf{X} = [x_i^\top] \in \mathbb{R}^{n \times d}$ and $\mathbf{Y} = [y_j^\top] \in \mathbb{R}^{m \times d}$, we consider an EM-like optimization scheme to calculate the empirically PRHCP distance, i.e., we optimize the transport plan \mathbf{P} and the $d \times q$ orthogonal matrix \mathbf{E} alternately and iteratively. Details for calculating PRHCP distance are summarized in Algorithm 2. This algorithm is similar to the one for calculating the subspace robust Wasserstein distance in [38], except that the transport plan is calculated by the HCP distance. As we observed in numerical experiments, Algorithm 2 performs well for high-dimensional cases and is robust to noise. Theoretical justification for these observations is left for future work.

Computational cost. For brevity, we consider the case that $n = m > d > q$. Step 3(a) requires $O(n \log(n)d)$ time, as discussed in the last section. Recall that there are at most $(n+m)$ nonzero elements in \mathbf{P} , and thus Step 3(b) requires only $O(n+m)d^2$ time. The cost for Step 3(c) involves $O(d^2q)$ for $\mathbf{U}\mathbf{U}^\top$ and $O(d^3)$ for solving the eigen-decomposition problem, respectively. Thus, the overall complexity of Algorithm 2 is $O(n \log(n)dL + nd^2L)$, where L is the number of iterations.

The proofs of above Theorems and their corollaries are given in Appendix.

Algorithm 2 Computation of PRHCP distance

- 1: **Input:** $(\mathbf{X} = \{x_i\}_{i=1}^n, \mathbf{a}), (\mathbf{Y} = \{y_j\}_{j=1}^m, \mathbf{b}), k, q$
 - 2: Initialize $\mathbf{U} = \Omega = \mathbf{I}_d, t = 0, \tau = 1$
 - 3: **While not converge**
 - a) $\mathbf{P} \leftarrow \text{Algorithm 1}[(\{\mathbf{U}^\top x_i\}_{i=1}^n, \mathbf{a}), (\{\mathbf{U}^\top y_j\}_{j=1}^m, \mathbf{b}), k]$
 $O((n \log(n) + m \log(m))d)$
 - b) $\mathbf{U} \in \mathbb{R}^{d \times q} \leftarrow \text{top } q \text{ singular vectors of the matrix } (\mathbf{X} - \text{diag}(\mathbf{a}^{-1})\mathbf{P}\mathbf{Y}) \text{ with weight } \mathbf{a}$
 $O((n + m)d^2)$
 - c) $\Omega \leftarrow (1 - \tau)\Omega + \tau\mathbf{U}\mathbf{U}^\top$, and then $\mathbf{U} \leftarrow \text{top } q \text{ eigenvectors of } \Omega$
 $O(d^2q + d^3)$
 - d) $t \leftarrow t + 1, \tau \leftarrow 2/(2 + t)$
 - 4: **Output:** The coupling \mathbf{P}_r and $\text{PRHCP}_{p,q} = (\sum_{(i,j) \in \{(i,j) | P_{ij} \neq 0\}} \|\mathbf{U}^\top x_i - \mathbf{U}^\top y_j\|_p^p P_{ij})^{1/p}$
-

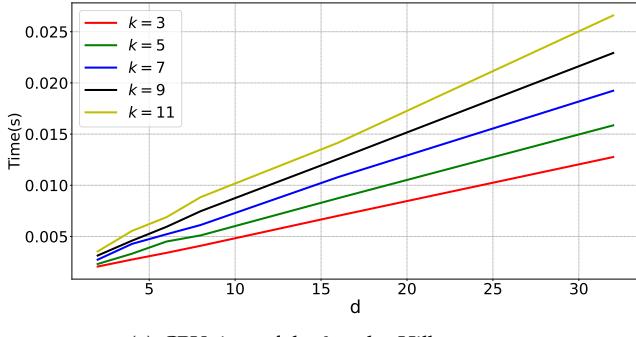
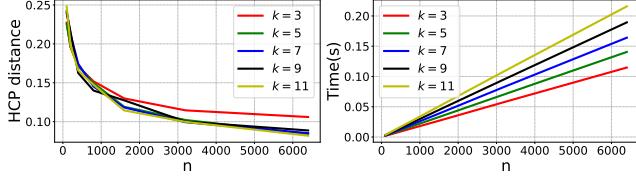
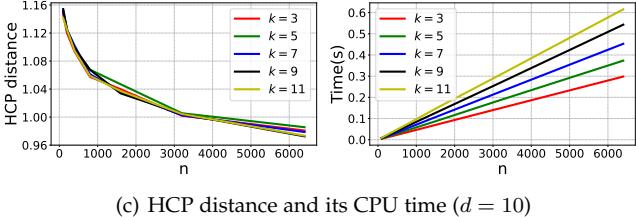
(a) CPU time of the k -order Hilbert curve(b) HCP distance and its CPU time ($d = 2$)

Fig. 4. (a) CPU time for generating the k -order Hilbert curve versus d when $n = 100$. (b) Left: HCP distance versus n when $d = 2$. Right: CPU time for generating the k -order Hilbert curve versus n when $d = 2$. (c) Left: HCP distance versus n when $d = 10$. Right: CPU time for generating the k -order Hilbert curve versus n when $d = 10$.

5 EXPERIMENTS

To demonstrate the feasibility and efficiency of our HCP distance and its variants, we conducted extensive numerical experiments and compared them with the mainstream competitors, including maximum mean discrepancy (MMD), Wasserstein distance, Sinkhorn distance [14], SW distance [15], max-SW distance [21], GSW distance [16], TSW distance [17], and PRW distance [39]. For all the distances, we considered the Euclidean cost, i.e., $p = 2$. We use k -order Hilbert curves with $k = 5 \log(n)$. We set the dimension for the intrinsic space as $q = 2$ for PRW, IPRHCP, and PRHCP.

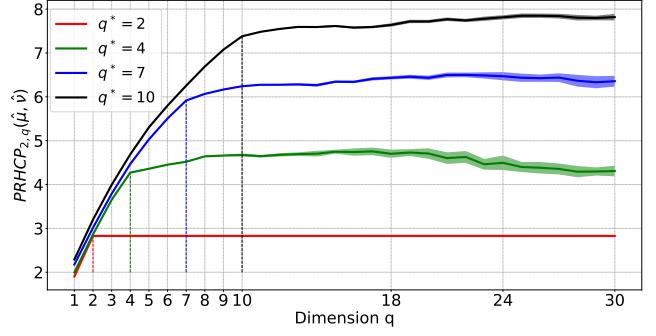


Fig. 5. $\text{PRHCP}_{2,q}(\mu_n, \nu_n)$ versus the dimension q for $q^* = 2, 4, 7, 10$.

All experiments are implemented by an AMD 3600 CPU and an RTX 1080Ti GPU. For each experiment, we replicate it 100 times and record the average performance.

5.1 Analytic experiments on synthetic data

5.1.1 Robustness and efficiency analysis

The performance of our HCP distance is mainly determined by three factors: (1) the order of the Hilbert curve; (2) the dimension of sample; and (3) the number of samples. To demonstrate the robustness and efficiency of our HCP distance, we test it on synthetic data and analyze the influences of the above three factors.

Specifically, we generate two sample sets of size n from the uniform distribution on the unit hypercube $[0, 1]^d$ and we calculate the HCP distance between these two sample sets. When calculating the HCP distance, the k -order Hilbert curve is applied. The results in Fig. 4(a) indicate the computational cost for generating the k -order Hilbert curve is linear to d . Figs. 4(b) and 4(c) show the average HCP distances and the average CPU time for generating the k -order Hilbert curve versus different n 's when $d = 2$ or 10 , respectively. From these two figures, we observe that the HCP distance is not sensitive to the choice of k , as long as k is not too small (i.e., $k > 3$). We also observe that the computational cost for generating the k -order Hilbert curve is linear to n .

For the variant of our HCP distance, i.e., the PRHCP distance, one more factor should be considered — the dimension of subspace. Ideally, this distance should be robust to the setting of q as long as q is equal to or larger than the dimension of the effective subspaces.

To demonstrate their robustness to q , we follow the settings in [38; 39], considering a uniform distribution $\mu = \mathcal{U}([-1, 1])^d$ and its pushforward under a map T , i.e., $\nu = T_\# \mu$. Here, the map $T(x) = x + 2 \text{sign}(x) \odot (\sum_{i=1}^{q^*} e_i)$, where sign is taken elementwise, $q^* = 2, 4, 7, 10$, and (e_1, \dots, e_d) is the canonical basis of \mathbb{R}^d . Obviously, the map T splits the hypercube into four different hyper-rectangles, and the dimension of the effective subspaces equals to q^* .

Setting $d = 50$ and $n = 100$, we calculate the PRHCP distance under different q 's. Fig. 5 shows the PRHCP distance increases rapidly when $q < q^*$ and tends to be stable and consistent when $q \geq q^*$. Such an observation indicates PRHCP can dig out useful subspace information effectively.

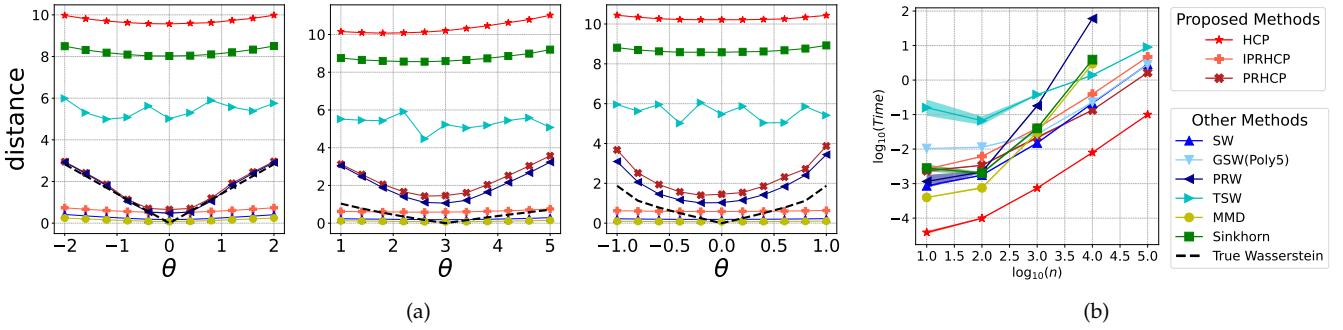


Fig. 6. The comparison for various metrics. (a) Distances versus different θ . (b) CPU time versus different n .

5.1.2 Comparisons in high-dimensional scenarios

As shown in Fig. 1, our HCP distance provides an effective and efficient surrogate of Wasserstein distance for 2D data. Here, we further compare various metrics on approximating Wasserstein distance for high-dimensional data. In particular, let $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ be i.i.d. samples generated from two Gaussian distributions, i.e., $\mathcal{N}_d(\mathbf{0}_d, \Sigma_X)$ and $\mathcal{N}_d(\mathbf{0}_d, \Sigma_Y)$, respectively. We consider three different settings as follows.

- 1) $\mu_X = \mathbf{0}_d$, $\mu_Y = (\theta, \theta, 0, \dots, 0)^\top$, $\Sigma_X = \Sigma_Y = \mathbf{I}_d$.
- 2) $\mu_X = \text{diag}(3\mathbf{I}_2, \mathbf{I}_{d-2})$, $\mu_Y = \text{diag}(\theta\mathbf{I}_2, \mathbf{I}_{d-2})$.
- 3) $\mu_X = \text{diag}(3\mathbf{I}_2, \mathbf{I}_{d-2})$, $\mu_Y = \text{diag}(\theta\mathbf{I}_2 + 3\theta\mathbf{B}_2, \mathbf{I}_{d-2})$.

where \mathbf{I}_d and \mathbf{B}_d are identity and backward identity matrices with size $(d \times d)$, respectively. In each of the three settings, the distance between the two distributions is controlled by a hyperparameter θ .

We set $n = 200$, $d = 50$. Given different θ 's, we generate different samples and calculate the distance between the two sample sets under different metrics. Fig. 6(a) shows the averaged distance in 100 trials. Taking the true Wasserstein distance between the two Gaussian densities as a benchmark, $W_2(\mathcal{N}_d(\mathbf{0}_d, \Sigma_X), \mathcal{N}_d(\mathbf{0}_d, \Sigma_Y)) = \text{tr}(\Sigma_X + \Sigma_Y - 2(\Sigma_X^{\frac{1}{2}} \Sigma_Y \Sigma_X^{\frac{1}{2}})^{\frac{1}{2}})$, we observe that most of the metrics, including our HCP distance, suffer from the curse-of-dimensionality or lack of robustness to noise, i.e., their distances are not sensitive to the parameter θ . Among these metrics, the PRW distance and our PRHCP distance are the only two that provide reasonable distances — they perform similarly as the true Wasserstein distance. In other words, although the HCP distance suffers from the curse-of-dimensionality, this problem can be mitigated by combining the HCP distance with the subspace projection strategy, leading to the PRHCP distance. Besides the comparison on the effectiveness, we also compare the CPU time for different metrics. Fig. 6(b) shows the CPU time (in seconds) versus different n 's. The time for our methods, including the HCP distance and its variants, is approximately linear to n . Compared to other metrics, our HCP requires significantly less time than all the competitors, and its two variants are at least comparable to other distances in runtime. Especially, our PRHCP distance works as well as the PRW distance does in high-dimensional scenarios, but its runtime is much less than the PRW distance's runtime, which demonstrates its superiority on both effectiveness and efficiency.

Additionally, we consider a synthetic example to demonstrate the empirical sample complexity of the proposed distances. We generate two samples of size n from the standard

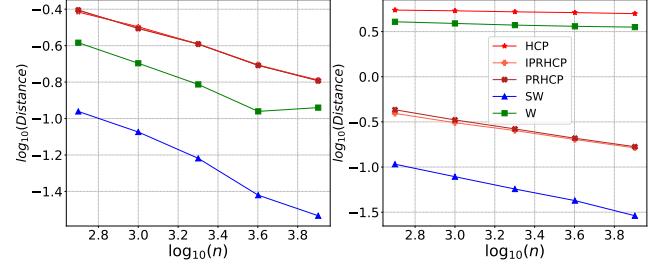


Fig. 7. Comparison for sample complexity. Left: $d = 2$. Right: $d = 20$. Each curve represents a distance versus n .

d -dimensional Gaussian distributions and we calculate the distances between these two samples w.r.t. different distance metrics. Fig. 7 shows the average distances versus n for $d = 2$ and 20, respectively. We observe that when $d = 20$, the Wasserstein distance and the HCP distance converge slowly as expected, while SW, IPRHCP, and PRHCP converge much faster. In the aspect of the empirical sample complexity, the slope of the curves indicates that our HCP distance is comparable to the Wasserstein distance, and our IPRHCP and PRHCP distances are comparable to the SW distance.

5.2 Approximation of Wasserstein flow

5.2.1 Comparison on synthetic data

Following the experiment in [16], we consider the problem $\min_{\mu} W_2(\mu, \nu)$, where ν is a fixed target distribution, and μ is the source distribution initialized as $\mu_0 = \mathcal{N}(0, 1)$ and updated iteratively via $\partial_t \mu_t = -\nabla W_2(\mu_t, \nu)$. We consider four different distributions for the target ν , i.e., *Circle*, *Swiss Roll*, *25-Gaussian*, and *Puma*, and approximate the Wasserstein distance W_2 by SW, max-SW, GSW, max-GSW, and HCP. Each method applies one projection per iteration and sets the learning rate to be 0.01. The experiments are replicated one hundred times, and we record the averaged 2-Wasserstein distance between μ_t and ν at each iteration. The comparison for the methods on their convergence curves and the snapshots of their learning results when $t = 150$ are shown in Fig. 8(a). We can find that applying HCP helps to accelerate the learning process and leads to better results.

Fig. 8(a) shows that using SW or its variants as the loss function may lead to slow convergence. Taking the *25-Gaussian* case as an example, we provide an intuitive explanation for this phenomenon. In particular, we illustrate

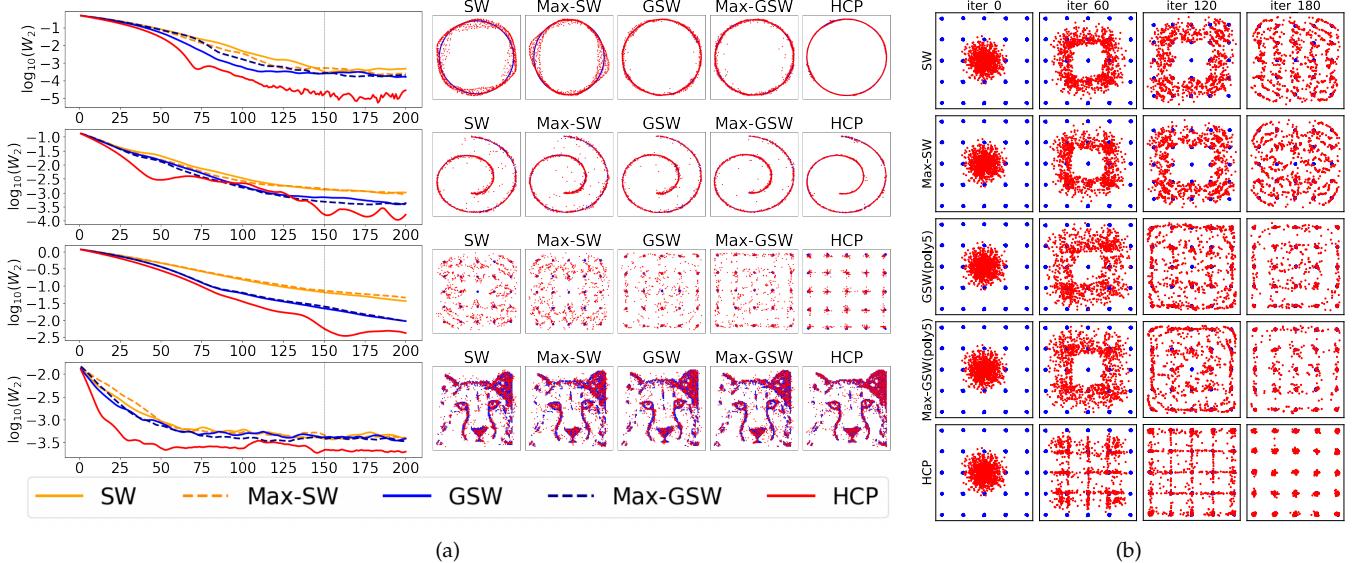


Fig. 8. (a) Left: Log 2-Wasserstein distance between the source and target distributions versus the number of iterations t . Right: A snapshot when $t = 150$. (b) Iterations of different distances based flow.

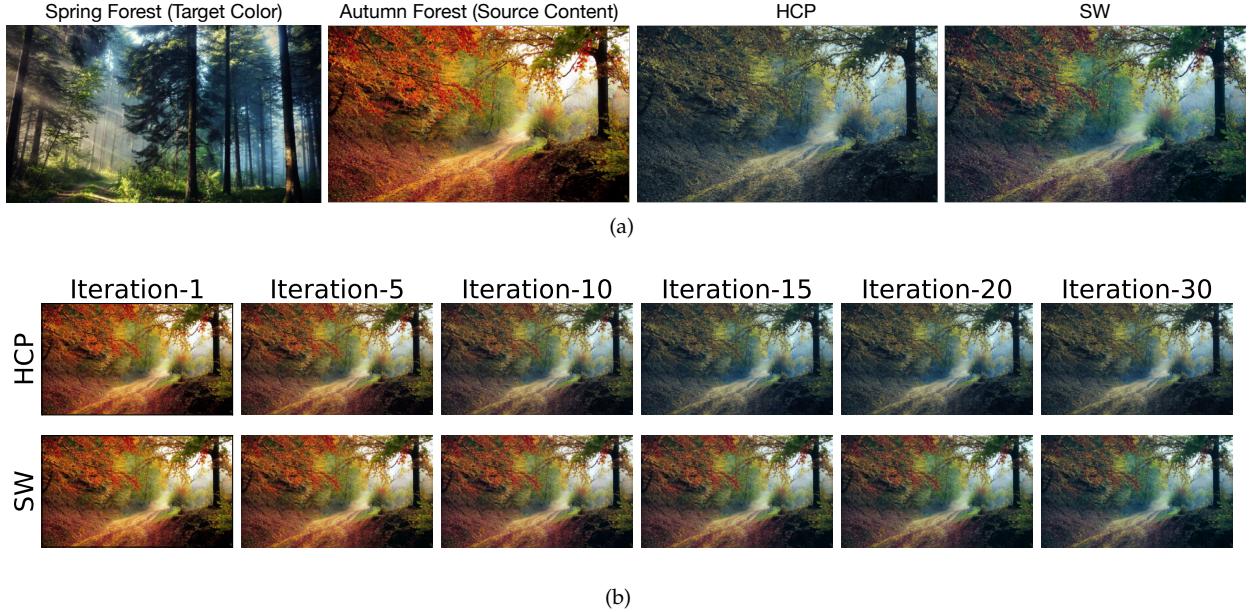


Fig. 9. (a) The images from left to right are the image with the target color, the image with the source content, and the color transfer results achieved based on our HCP distance and the SW distance, respectively. (b) The first row is iterations of color transfer based on our HCP distance. The second row is iterations of color transfer based on the SW distance.

the iterations of SW, Max-SW, GSW, Max-GSW and HCP in Fig. 8(b). We observe that the flow w.r.t. SW and its variants go through 2 processes: firstly, red points spread out without covering the central Gaussian; secondly, they cover the central Gaussian slowly. Such an observation indicates linear projection fails to preserve high-dimensional data structure, especially when the data are multi-modal, and thus resulting in slow convergence. The proposed HCP distance, on the contrary, utilizes a Hilbert curve to preserve the structure of high-dimensional data and thus leads to faster convergence.

5.2.2 Color transfer for images

Besides testing on synthetic data, we consider the real-world color transfer task. As shown in Fig. 9(a), we transfer the color of a *Spring Forest* image to an *Autumn Forest* image. Each image is represented as nearly two million pixels in the RGB space ($d = 3$). Considering the large sample size, we use SW distance and HCP distance to approximate the Wasserstein flow, with the same learning hyperparameters. The comparison of the methods on their color transfer results and iterations are shown in Fig. 9. We can find that applying the HCP distance helps to accelerate the learning process. Quantitatively, it takes 496.7 seconds for the SW-based method and 57.3 seconds for our HCP-based method.

TABLE 1
Comparisons on 3D point cloud classification

Method	Accuracy(%)					CPU time(s)				
	n=100	n=200	n=500	n=1000	n=2000	n=100	n=200	n=500	n=1000	n=2000
HCP	73.3 \pm 1.3	79.2 \pm 2.8	81.8 \pm 0.2	81.0 \pm 0.1	82.3 \pm 0.7	17.9	34.3	90.4	186.7	374.5
SW	71.5 \pm 2.2	77.0 \pm 1.0	79.2 \pm 0.5	79.3 \pm 0.7	80.5 \pm 1.0	123.9	202.1	410.5	830.8	1808.7
TSW	72.7 \pm 3.2	75.0 \pm 2.5	77.0 \pm 2.2	77.7 \pm 1.5	78.0 \pm 1.0	120.7	137.4	165.8	217.2	282.1
GSW(Poly5)	68.3 \pm 2.0	73.2 \pm 0.5	76.0 \pm 0.7	76.8 \pm 1.2	77.7 \pm 0.3	839.7	940.6	1228.8	1782.4	2970.2
MMD	66.8 \pm 4.2	71.8 \pm 0.2	74.0 \pm 0.7	/	/	153.6	385.7	1785.2	/	/
Sinkhorn	77.0 \pm 2.7	80.8 \pm 0.5	81.5 \pm 0.8	/	/	768.7	2305.0	6184.6	/	/

* “/” means that we fail to get a result in 10,000 seconds.

5.3 Data classification

5.3.1 3D point cloud classification

For low-dimensional data like 3D points, our HCP distance is superior to other distances in their classification tasks. We consider the ModelNet10 dataset [73] that contains around 5,000 CAD objects from 10 categories. For each category, we randomly sample 50 objects for training and 30 object objects for testing. Following the work in [74], we randomly sample $n = 100, 200, 500, 1000, 2000$ points per object to get 3D point cloud data. We calculate the pairwise distance between the point clouds w.r.t. different distance metrics and then use the K-NN algorithm ($n_{neighbors} = 5$) to evaluate the classification accuracy on the testing set. We used the RBF kernel for MMD, and we set the number of slices $n_s = 10$ for SW, $n_s = 10, T = 7, \kappa = 4$ for TSW. Here, T is the predefined deepest level of the tree, n_s is the number of slices and κ is the number of clusters. Table 1 summarizes the averaged performance of each metric in 10 trials. Our HCP outperforms other distances on accuracy and requires the least amount of time.

5.3.2 Document classification

As a typical high-dimensional data classification problem, document classification can be achieved by comparing the Wasserstein distance between two documents’ word embedding sets, as the Word Mover distance [1] does. Our PRHCP distance provides an efficient surrogate of the Wasserstein distance in this problem, which is demonstrated by the following experiment. Following the preprocessing used in [1], we obtain 3,000 documents belonging to three categories from the TWITTER dataset, in which each document is represented as a set of 300-dimensional word embeddings derived by the pre-trained *word2vec* model [75]. We randomly split the dataset into 80% for training and 20% for testing. Similar to the above point cloud classification experiment, we use the K-NN algorithm ($n_{neighbors} = 10$) based on different metrics and evaluate the averaged learning results in 10 trials. In this experiment, we set the number of slices $n_s = 20$ for SW, $n_s = 10, T = 7, \kappa = 4$ for TSW. For PRHCP, we first find the 10-dimensional subspace based on the training data by Algorithm 2 and project testing data to the subspace. Table 2 shows that our PRHCP distance outperforms other distances on classification accuracy, and its runtime is comparable to TSW.

TABLE 2
Comparisons on document classification

Method	Accuracy(%)	CPU time(s)
PRHCP	74.6 \pm 0.9	669.7
TSW	71.2 \pm 0.6	287.7
Sinkhorn	70.0 \pm 0.4	10106.7
SW	68.6 \pm 1.1	2789.7

5.4 Generative modeling

The proposed distances help us to design new members of Wasserstein autoencoder (WAE) [9]. In particular, when training autoencoders, we leverage HCP, IPRHCP, and PRHCP to penalize the distance between the latent prior distribution and the expected posterior distribution, which leads to three different generative models, denoted as HCP-AE, IPRHCP-AE, and PRHCP-AE. We test these three models in image generation tasks and compare them with the original Wasserstein autoencoder (WAE) [9] and the well-known sliced Wasserstein autoencoder (SWAE) [22].

5.4.1 HCP-based autoencoders

We first test the capability of HCP-AE in shaping the low-dimensional latent space of the encoder. We train an HCP-AE to encode the MNIST dataset [76] to a two-dimensional latent space (for the sake of visualization), in which both the autoencoding architecture and the hyperparameter setting are the same as those in [22]. A simple autoencoder with mirrored classic deep convolutional neural networks with 2D average poolings, Leaky-ReLu activation functions, and upsampling layers in the decoder is used. The batch size is 500 and the number of projections for SWAE is 40.

To evaluate the performance, we randomly selected a sample of size 1,000 from the encoded test data points (blue points in Fig. 10(b)) and a random sample from the target prior distribution in the latent space (red points in Fig. 10(b)). We observed that our HCP-AE converges much faster than other methods in the latent space. Moreover, the SW distances versus the number of epochs w.r.t. the image space and the latent space are shown in Fig. 10(a). We observed that though these three methods perform similarly in the image space, our HCP-AE converges much faster in the latent space. Fig. 10(c) visualizes the samples from two different prior distributions in the latent space, the encoded data samples via HCP-AE, and their generated images. The latent

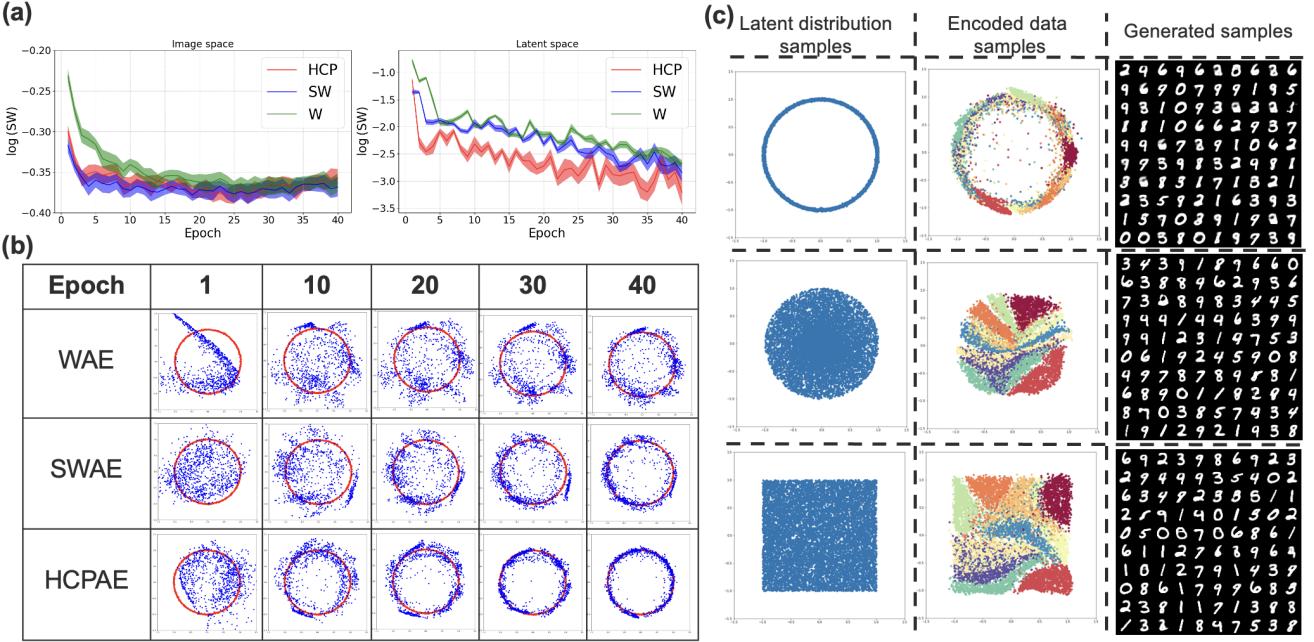


Fig. 10. (a): SW distances between the target sample and the encoded testing sample w.r.t. the image space (left) and the latent space (right); (b) Visualization of these two sample in the latent space during training; (c) Visualization of the encoded samples and the generated images.

codes indeed obey the prior distributions, which reflects the clustering structure of the digits. Accordingly, the learned models are able to generate high quality digit images.

5.4.2 IPRHCP and PRHCP-based autoencoders

Secondly, we test the feasibility of IPRHCP-AE and PRHCP-AE in the cases with high-dimensional latent space. For fairness, all the autoencoders have the same DCGAN-style architecture [77] and hyperparameters: the learning rate is 0.001; the optimizer is Adam [78] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$; the number of epochs is 50; the batch size is 100; the weight of regularizer γ is 1; the dimension of latent code is 8 for MNIST and 64 for CelebA; the number of random projections is 50. All the autoencoders use Euclidean distance as the distance between samples, which means the reconstruction loss is the mean-square error (MSE). We compare the proposed methods with the baselines on *i*) the reconstruction loss on testing samples; *ii*) the Fréchet Inception Distance (FID) [79] between 10,000 testing samples and 10,000 randomly generated samples. Table 3 lists the main differences between IPRHCP-AE, PRHCP-AE and these baselines. Among these autoencoders, our IPRHCP-AE and PRHCP-AE are comparable to the considered alternatives on both testing reconstruction loss and FID score. Some image generation and interpolation results achieved by our methods are shown in Fig. 11.

6 CONCLUSION

In this work, we proposed a novel metric for distribution comparison, named Hilbert curve projection (HCP) distance. Thanks to the locality-preserving property of the Hilbert curve projection, the HCP distance enjoys several advantages over the Wasserstein and SW distance. Furthermore, we

TABLE 3
Comparisons for various methods on learning image generators

Method	MNIST		CelebA	
	Rec. loss	FID	Rec. loss	FID
WAE	11.30	54.61 ± 0.16	68.94	58.12 ± 0.73
SWAE	13.68	42.96 ± 0.53	68.57	84.52 ± 0.44
IPRHCP-AE	11.72	40.03 ± 0.13	69.40	56.00 ± 0.08
PRHCP-AE	10.07	42.87 ± 0.46	66.65	67.82 ± 0.21

develop two variants of the HCP distance using (learnable) subspace projections to mitigate the curse-of-dimensionality.

Limitations and future work. Currently, HCP distance still suffers from some limitations. Like the Wasserstein distance, the HCP distance may not be robust to outliers. To address this problem, we could follow the methods in [80; 81; 82] by relaxing marginal constraints through penalty functions such as Kullback-Leibler divergence, total variation distance, and χ^2 divergence. Another possible solution is to consider partial OT methods instead of sorting in Step 3 of Algorithm 1. Besides, HCP distance could not quantify the discrepancy between two measures with different masses. We could follow the idea of (sliced) unbalanced optimal transport [83; 84; 85] by considering unbalanced OT methods instead of sorting in Step 3 of Algorithm 1. We left these directions for our future work. In addition, we plan to apply these new metrics to more learning problems and extend them to Gromov-Wasserstein distance [86], multi-marginal optimal transport [87; 88], and barycenter problems [89; 90]. Additionally, we will explore the theoretical results for other formulations of the Hilbert curve, such as the adaptive Hilbert curve, which works well in practice. And there is much literature on Hilbert sort, such as parallel Hilbert



Fig. 11. The performance of IPRHCP-AE and PRHCP-AE on face generation and interpolation.

sort [91] and online Hilbert sort [69; 92], which may be extended.

ACKNOWLEDGMENT

All authors contributed equally and the order of authors' names is alphabetical. The authors would like to acknowledge the support from National Natural Science Foundation of China Grant No.62106271, No.12101606, No.12001042, and Renmin University of China research fund program for young scholars, and Beijing Institute of Technology research fund program for young scholars.

APPENDIX A THE PROOFS OF THEOREMS AND COROLLARIES

A.1 Preliminaries

The Hilbert curve enjoys three properties, see [57; 58] for details.

- 1) Let λ_d be the d -dimensional Lebesgue measure. For any measurable set $\Omega \subseteq [0, 1]$, one has $\lambda_1(\Omega) = \lambda_d(H(\Omega))$.
- 2) If a random variable Z follows the uniform distribution on $[0, 1]$, then $H(Z)$ follows the uniform distribution on $[0, 1]^d$. This is to say, one has $\int_{[0,1]^d} f(z)dz = \int_0^1 f[H(z)]dz$.
- 3) For any $x, y \in [0, 1]$, one has $\|H(x) - H(y)\|_2 \leq 2\sqrt{d+3}|x-y|^{1/d}$.

The Hölder continuous property also holds for H_μ . Considering that H_μ is a linear stretching of H and using Cauchy-Schwarz inequality, we have $\|H_\mu(x) - H_\mu(y)\|_2 \leq C_\mu \|H(x) - H(y)\|_2 \leq 2C_\mu \sqrt{d+3}|x-y|^{1/d}$ for any $x, y \in [0, 1]$. And by equivalence of vector norms,

we have $\|H_\mu(x) - H_\mu(y)\|_p \leq C_p \|H_\mu(x) - H_\mu(y)\|_2 \leq 2C_\mu C_p \sqrt{d+3}|x-y|^{1/d}$, where C_p, C_μ are two constants.

Before presenting the main statistical theoretical results of HCP distance, we first provide an essential lemma.

Lemma 1. *If a random variable Z follows the probability distribution function g_μ , then $H_\mu(Z)$ follows the probability measure μ .*

Proof of Lemma 1. By the definition of g_μ , we can easily know that $Pr(Z \in A) = \inf\{\mu(A') : A' \text{ is a Borel set in } \mathbb{R}^d, H_\mu(A \setminus \mathcal{N}) \subset A'\}$ for any Borel measurable set $A \subset \mathbb{R}$. And for any measurable set $B \subset \mathbb{R}^d$, $Pr(H_\mu(Z) \in B) = Pr(Z \in H_\mu^{-1}(B) \setminus \mathcal{N}) = \inf\{\mu(A') : A' \text{ is a Borel set in } \mathbb{R}^d, H_\mu(H_\mu^{-1}(B) \setminus \mathcal{N}) \subset A'\} = \inf\{\mu(A') : A' \text{ is a Borel set in } \mathbb{R}^d, B \subset A'\} = \mu(B)$. \square

A.2 The proofs related to HCP distance

It is clear that HCP distance finds a transport plan between two probability measures by Hilbert sorting. Thus, it obviously gives an upper bound of Wasserstein distance. Below, we clarify it theoretically and prove HCP distance is a well-defined distance.

Proof of Theorem 1. Here, we first borrow the Corollary 3.1 proven below. It tells us that $\mathbb{E} \frac{1}{n} \sum_{i=1}^n \|x_{(i)*} - y_{(i)*}\|_p^p \rightarrow HCP_p(\mu, \nu)$, as $n \rightarrow \infty$. Additionally, we know that $\mathbb{E} \inf_{\sigma} \frac{1}{n} \sum_{i=1}^n \|x_i - y_{\sigma(i)}\|_p^p \rightarrow W_p^p(\mu, \nu)$, as $n \rightarrow \infty$. Clearly, $\inf_{\sigma} \frac{1}{n} \sum_{i=1}^n \|x_i - y_{\sigma(i)}\|_p^p \leq \frac{1}{n} \sum_{i=1}^n \|x_{(i)*} - y_{(i)*}\|_p^p$. Hence, we have $W_p(\mu, \nu) \leq HCP_p(\mu, \nu)$.

By the definition of HCP, it's clear that $\text{HCP}_p(\mu, \nu) \geq 0$ and $\text{HCP}_p(\mu, \nu) = \text{HCP}_p(\nu, \mu)$. Because Wasserstein distance is a well-defined metric, we can easily get $\text{HCP}_p(\mu, \nu) = 0$ if and only if $\mu = \nu$. Next, we prove the triangle inequality.

$$\begin{aligned} & \text{HCP}_p(\mu, \nu) \\ &= \left(\int_0^1 \|H_\mu(g_\mu^{-1}(t)) - H_\nu(g_\nu^{-1}(t))\|_p^p dt \right)^{1/p} \\ &= \left(\int_0^1 \theta_{\mu, \nu}(t)^p dt \right)^{1/p} \\ &\leq \left(\int_0^1 \theta_{\mu, \tau}(t)^p dt \right)^{1/p} + \left(\int_0^1 |\theta_{\mu, \nu}(t) - \theta_{\mu, \tau}(t)|^p dt \right)^{1/p} \\ &\leq \left(\int_0^1 \theta_{\mu, \tau}(t)^p dt \right)^{1/p} + \left(\int_0^1 \theta_{\nu, \tau}(t)^p dt \right)^{1/p} \\ &= \text{HCP}_p(\mu, \tau) + \text{HCP}_p(\nu, \tau), \end{aligned}$$

where first inequality uses Minkowski inequality and the second uses triangle inequality. \square

Proof of Theorem 2. From the proof in Theorem 3, we have for some $\alpha > 0$,

$$\text{HCP}_p(\mu_n, \mu) \leq \left(\int_0^1 |g_{\mu_n}^{-1}(t) - g_\mu^{-1}(t)| dt \right)^\alpha = W_1(\tilde{\mu}_n, \tilde{\mu})^\alpha,$$

where the probability measure corresponding to g_{μ_n} is $\tilde{\mu}_n$ and the probability measure corresponding to g_μ is $\tilde{\mu}$.

Since $TV(\mu_n, \mu) \rightarrow 0$, we have $\forall \epsilon$, there exists a N_0 , $\sup_A |\mu_n(A) - \mu(A)| < \epsilon$ when $n > N_0$. For any measurable set $B \subset [0, 1]$, we have $|\tilde{\mu}_n(B) - \tilde{\mu}(B)| < \epsilon$ when $n > N_0$. Hence, we know $W_1(\tilde{\mu}_n, \tilde{\mu}) \rightarrow 0$, and the proof is done. \square

Proof of Theorem 3. Let $\{y_i\}_{i=1}^n$ be an i.i.d sample, which is generated from probability distribution function g_μ . By Lemma 1, we could replace $\{x_i\}_{i=1}^n$ with $\{H_\mu(y_i)\}_{i=1}^n$, and thus, replace $\hat{g}_{\mu_n}(t)$ with $g_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{y_i \leq t\}}$. Hence, to study the convergence of $\overline{\text{HCP}}_p(\mu, \mu_n) = (\int_0^1 \|H_\mu(g_\mu^{-1}(t)) - H_\mu(g_n^{-1}(t))\|_p^p dt)^{\frac{1}{p}}$, we only need to study $(\int_0^1 \|H_\mu(g_\mu^{-1}(t)) - H_\mu(g_n^{-1}(t))\|_p^p dt)^{\frac{1}{p}}$.

We know that

$$\begin{aligned} & \left(\int_0^1 \|H_\mu(g_\mu^{-1}(t)) - H_\mu(g_n^{-1}(t))\|_p^p dt \right)^{\frac{1}{p}} \\ &\lesssim \left(\int_0^1 |g_n^{-1}(t) - g_\mu^{-1}(t)|^{\frac{p}{d}} dt \right)^{\frac{1}{p}}, \end{aligned}$$

and we have when $p \geq d$

$$\left(\int_0^1 |g_n^{-1}(t) - g_\mu^{-1}(t)|^{\frac{p}{d}} dt \right)^{\frac{1}{p}} \leq \left(\int_0^1 |g_n^{-1}(t) - g_\mu^{-1}(t)| dt \right)^{\frac{1}{d}}.$$

When $p < d$, by Jensen's inequality, we have

$$\left(\int_0^1 |g_n^{-1}(t) - g_\mu^{-1}(t)|^{\frac{p}{d}} dt \right)^{\frac{1}{p}} \leq \left(\int_0^1 |g_n^{-1}(t) - g_\mu^{-1}(t)| dt \right)^{\frac{1}{d}}.$$

However, $\int_0^1 |g_n^{-1}(t) - g_\mu^{-1}(t)| dt$ is the 1-Wasserstein distance. Considering g_μ has bounded support, we know that $\int_0^1 |g_n^{-1}(t) - g_\mu^{-1}(t)| dt \rightarrow 0$ almost surely. Thus, we know that $\overline{\text{HCP}}_p(\mu, \mu_n) \rightarrow 0$ almost surely.

Furthermore, we know that

$$\begin{aligned} \mathbb{E} \overline{\text{HCP}}_p(\mu, \mu_n) &= \mathbb{E} \left(\int_0^1 \|H_\mu(g_n^{-1}(t)) - H_\mu(g_\mu^{-1}(t))\|_p^p dt \right)^{\frac{1}{p}} \\ &\lesssim \mathbb{E} \left(\int_0^1 |g_n^{-1}(t) - g_\mu^{-1}(t)|^{\frac{p}{d}} dt \right)^{\frac{1}{p}} \\ &\leq \left(\mathbb{E} \int_0^1 |g_n^{-1}(t) - g_\mu^{-1}(t)|^{\frac{p}{d}} dt \right)^{\frac{1}{p}}. \end{aligned}$$

If $p \geq d$, we know that

$$\begin{aligned} \mathbb{E} \overline{\text{HCP}}_p(\mu, \mu_n) &\lesssim \left(\mathbb{E} \int_0^1 |g_n^{-1}(t) - g_\mu^{-1}(t)|^{\frac{p}{d}} dt \right)^{\frac{1}{p}} \\ &\lesssim \left(\mathbb{E} \int_0^1 |g_n^{-1}(t) - g_\mu^{-1}(t)| dt \right)^{\frac{1}{p}} \\ &\lesssim O(n^{-\frac{1}{2p}}). \end{aligned}$$

The last equality is by the upper bound of 1-Wasserstein distance [63].

If $p < d$, we know that

$$\begin{aligned} \mathbb{E} \overline{\text{HCP}}_p(\mu, \mu_n) &\lesssim \left(\mathbb{E} \int_0^1 |g_n^{-1}(t) - g_\mu^{-1}(t)|^{\frac{p}{d}} dt \right)^{\frac{1}{p}} \\ &\lesssim \left(\mathbb{E} \left(\int_0^1 |g_n^{-1}(t) - g_\mu^{-1}(t)| dt \right)^{\frac{p}{d}} \right)^{\frac{1}{p}} \\ &\lesssim \left(\mathbb{E} \left(\int_0^1 |g_n^{-1}(t) - g_\mu^{-1}(t)| dt \right) \right)^{\frac{1}{d}} \\ &\lesssim O(n^{-\frac{1}{2d}}). \end{aligned}$$

The second and third inequalities use Jensen's inequality and the last inequality is by the upper bound of 1-Wasserstein distance [63]. \square

Proof of Corollary 3.1. By definition, we know that $\overline{\text{HCP}}_p(\mu_n, \nu_n) = (\frac{1}{n} \sum_{i=1}^n \|x_{(i)*} - y_{(i)*}\|_p^p)^{\frac{1}{p}}$. By triangle inequality, we have

$$\begin{aligned} \overline{\text{HCP}}_p(\mu_n, \nu_n) - \text{HCP}_p(\mu, \nu) &\geq -\overline{\text{HCP}}_p(\mu, \mu_n) - \overline{\text{HCP}}_p(\nu, \nu_n), \\ \overline{\text{HCP}}_p(\mu_n, \nu_n) - \text{HCP}_p(\mu, \nu) &\leq \overline{\text{HCP}}_p(\mu, \mu_n) + \overline{\text{HCP}}_p(\nu, \nu_n). \end{aligned}$$

By Theorem 3, we can easily conclude the almost surely converge property.

By Theorem 3, we have

$$\overline{\text{HCP}}_p(\mu, \mu_n) \lesssim O(n^{-\frac{1}{2 \max\{p, d\}}}),$$

and

$$\overline{\text{HCP}}_p(\nu, \nu_n) \lesssim O(n^{-\frac{1}{2 \max\{p, d\}}}).$$

Thus, we have

$$\overline{\text{HCP}}_p(\mu_n, \nu_n) - \text{HCP}_p(\mu, \nu) \lesssim O(n^{-\frac{1}{2 \max\{p, d\}}}),$$

and

$$\overline{\text{HCP}}_p(\mu_n, \nu_n) - \text{HCP}_p(\mu, \nu) \gtrsim -O(n^{-\frac{1}{2 \max\{p, d\}}}).$$

Hence, we have finished the proof. \square

Corollary 3.1 can be easily extended to samples with different weights, which are omitted here.

Proof of Corollary 3.2. It can be similarly concluded like Corollary 3.1. \square

Proof of Corollary 3.3. By Section 3.3 in [63], we know that $\mathbb{E}W_p(\mu_n, \mu) \geq O(n^{-\frac{1}{2p}})$. From Theorem 3, we have proved it. \square

Proof of Corollary 3.4. Without loss of generality, we assume s_1, s_2, \dots, s_K are sorted in Hilbert order and $s_i \neq s_j$ for any $i \neq j$. Since there are finite supports, we easily know that $Pr(A^c) \leq \sum_{i=1}^K Pr(\{x_j \neq s_i, \forall j\}) = \sum_{i=1}^K (p_i)^n$ where $A = \{\tilde{\Omega}_\mu = \tilde{\Omega}_{\mu_n}\}$. Denote $\hat{p}_i = \frac{\sum_{j=1}^n \mathbf{1}_{\{s_i\}}(x_j)}{n}$ where $\mathbf{1}$ is the indicator function. Following the notation in Theorem 3, we have

$$\begin{aligned} \mathbb{E}HCP_p(\mu, \mu_n) &= \mathbb{E}[HCP_p(\mu, \mu_n)\mathbf{1}_A + HCP_p(\mu, \mu_n)\mathbf{1}_{A^c}] \\ &\lesssim \left(\mathbb{E} \int_0^1 |g_n^{-1}(t) - g_\mu^{-1}(t)|^{\frac{p}{d}} dt \right)^{\frac{1}{p}} + Pr(A^c) \\ &\lesssim \left(\mathbb{E}W_1 \left(\sum_{i=1}^K p_i \delta_{\{s_i\}}, \sum_{i=1}^K \hat{p}_i \delta_{\{s_i\}} \right) \right)^{\frac{1}{p}} + \sum_{i=1}^K (p_i)^n \\ &\lesssim O(n^{-1/2p}). \end{aligned}$$

The second inequality is directly from the discrete distribution and the last inequality is from [63].

We have proved the upper bound. The lower bound can be directly concluded by $W_p(\mu_n, \mu) \gtrsim n^{-1/2p}$; see [63]. \square

A.3 The proofs related to IPRHCP and PRHCP

Proof of Theorem 4. By definition and Theorem 1

$$\begin{aligned} \text{IPRHCP}_{p,q}(\mu, \nu) &= \left(\int_{\mathbb{S}_{d,q}} HCP_p^p(P_{\mathbf{E}\#}\mu, P_{\mathbf{E}\#}\nu) d\sigma(\mathbf{E}) \right)^{1/p} \\ &\geq \left(\int_{\mathbb{S}_{d,q}} W_p^p(P_{\mathbf{E}\#}\mu, P_{\mathbf{E}\#}\nu) d\sigma(\mathbf{E}) \right)^{1/p} \\ &= \text{IPRW}_{p,q}(\mu, \nu). \end{aligned}$$

By definition of IPRHCP, it is clear that $\text{IPRHCP}_{p,q}(\mu, \nu) \geq 0$ and $\text{IPRHCP}_{p,q}(\mu, \nu) = \text{IPRHCP}_{p,q}(\nu, \mu)$. Because $\text{IPRW}_{p,q}$ is a well-defined metric, we can easily get $\text{IPRHCP}_{p,q}(\mu, \nu) = 0$ if and only if $\mu = \nu$. Next, we prove the triangle inequality as follows,

$$\begin{aligned} \text{IPRHCP}_{p,q}(\mu, \nu) &= \left(\int_{\mathbb{S}_{d,q}} HCP_p^p(P_{\mathbf{E}\#}\mu, P_{\mathbf{E}\#}\nu) d\sigma(\mathbf{E}) \right)^{\frac{1}{p}} \\ &\leq \left(\int_{\mathbb{S}_{d,q}} [HCP_p(P_{\mathbf{E}\#}\mu, P_{\mathbf{E}\#\tau}) + \right. \\ &\quad \left. HCP_p(P_{\mathbf{E}\#}\nu, P_{\mathbf{E}\#\tau})]^p d\sigma(\mathbf{E}) \right)^{\frac{1}{p}} \\ &\leq \text{IPRHCP}_{p,q}(\mu, \tau) + \text{IPRHCP}_{p,q}(\nu, \tau), \end{aligned}$$

where the first inequality uses triangle inequality from Theorem 1 and the second uses Minkowski inequality. \square

Proof of Theorem 5. First, we introduce an inequality between W and SW distance, which has been proved in [93]. When $c(x, y) = \|x - y\|_p^p$ for $p \geq 2$, the following upper bound hold for the SW distance:

$$SW_p^p(\mu, \nu) \leq \alpha_{d,p} W_p^p(\mu, \nu),$$

where $\alpha_{d,p} = \frac{1}{d} \int_{\mathbb{S}_{d,1}} \|\theta\|_p^p d\theta \leq 1$.

Thus, we have:

$$\begin{aligned} &\text{IPRHCP}_{p,q}^p(\mu, \nu) \\ &= \int_{\mathbb{S}_{d,q}} HCP_p^p(P_{\mathbf{E}\#}\mu, P_{\mathbf{E}\#}\nu) d\sigma(\mathbf{E}) \\ &\geq \int_{\mathbb{S}_{d,q}} W_p^p(P_{\mathbf{E}\#}\mu, P_{\mathbf{E}\#}\nu) d\sigma(\mathbf{E}) \\ &\geq \int_{\mathbb{S}_{d,q}} \frac{1}{\alpha_{q,p}} SW_p^p(P_{\mathbf{E}\#}\mu, P_{\mathbf{E}\#}\nu) d\sigma(\mathbf{E}) \\ &= \frac{1}{\alpha_{q,p}} \int_{\mathbb{S}_{d,q}} \int_{\mathbb{S}_{q,1}} W_p^p(P_{\mathbf{v}\#}[P_{\mathbf{E}\#}\mu], P_{\mathbf{v}\#}[P_{\mathbf{E}\#}\nu]) d\sigma(\mathbf{v}) d\sigma(\mathbf{E}) \\ &= \frac{1}{\alpha_{q,p}} \int_{\mathbb{S}_{d,1}} W_p^p(P_{\mathbf{v}\#}\mu, P_{\mathbf{v}\#}\nu) d\sigma(\mathbf{v}) \\ &= \frac{1}{\alpha_{q,p}} SW_p^p(\mu, \nu), \end{aligned}$$

where the first inequality uses Theorem 1, the second inequality uses the above inequality and the last several equalities use the property of SW. \square

Proof of Corollary 5.1. Here, we take $q = 2$ as an example. It can be easily extended to the case $q > 2$. From the construction of 2-d Hilbert Curve, we know that if $x_1 < x_2$, then $\min\{H^{-1}((x_1, x_1))\} < \min\{H^{-1}((x_2, x_2))\}$. By $\mathbf{E}^\top \mathbf{E} = \mathbf{J}_2$, we know $\mathbf{E} = (\mathbf{v}, \mathbf{v})$ with $\|\mathbf{v}\|_2 = 1$. Hence, we can easily conclude that

$$HCP_p^p(P_{\mathbf{E}\#}\mu, P_{\mathbf{E}\#}\nu) = 2W_p^p(P_{\mathbf{v}\#}\mu, P_{\mathbf{v}\#}\nu).$$

Thus,

$$\begin{aligned} \text{IPRHCP}_{p,q}(\mu, \nu) &= \left(\int HCP_p^p(P_{\mathbf{E}\#}\mu, P_{\mathbf{E}\#}\nu) d\sigma(\mathbf{E}) \right)^{1/p} \\ &= \left(\int_{\mathbb{S}_{d,1}} 2W_p^p(P_{\mathbf{v}\#}\mu, P_{\mathbf{v}\#}\nu) d\sigma(\mathbf{v}) \right)^{1/p} \\ &= 2^{1/p} SW_p(\mu, \nu). \end{aligned}$$

\square

Proof of Theorem 6. From the proof of Theorem 3, we know that for any given subspace \mathbf{E} ,

$$\begin{aligned} &\mathbb{E} \frac{1}{n} \sum_{i=1}^n \int_{\frac{i-1}{n}}^{\frac{i}{n}} \|x_{\mathbf{E},(i)*} - H_{P_{\mathbf{E}\#}\mu}(g_{P_{\mathbf{E}\#}\mu}^{-1}(t))\|_p^p dt \\ &\leq C_1 \mathbb{E} \int_0^1 |g_n^{-1}(\mathbf{E})(t) - g_{P_{\mathbf{E}\#}\mu}^{-1}(t)|^{\frac{p}{d}} dt, \end{aligned}$$

where $g_n^{-1}(\mathbf{E})(t)$, $g_{P_{\mathbf{E}\#}\mu}^{-1}(t)$ are defined similar to Theorem 3 and where C_1 is a constant, not depending on \mathbf{E} .

Considering μ has bounded support, from [94], we know that for any given \mathbf{E} ,

$$\mathbb{E} \left(\int_0^1 |g_n^{-1}(\mathbf{E})(t) - g_{P_{\mathbf{E}\#}\mu}^{-1}(t)| dt \right) \leq \frac{C_2}{\sqrt{n}},$$

where C_2 is a constant, not depending on \mathbf{E} .

When $p \geq q$, we have

$$\begin{aligned} &\mathbb{E} \int_0^1 |g_n^{-1}(\mathbf{E})(t) - g_{P_{\mathbf{E}\#}\mu}^{-1}(t)|^{\frac{p}{d}} dt \\ &\leq \mathbb{E} \int_0^1 |g_n^{-1}(\mathbf{E})(t) - g_{P_{\mathbf{E}\#}\mu}^{-1}(t)| dt \leq \frac{C_2}{\sqrt{n}}. \end{aligned}$$

When $p < q$, by Jensen's inequality,

$$\begin{aligned} & \mathbb{E} \int_0^1 |g_{n,\mathbf{E}}^{-1}(t) - g_{P_{\mathbf{E}\#}\mu}^{-1}(t)|^{\frac{p}{q}} dt \\ & \leq \left(\mathbb{E} \int_0^1 |g_{n,\mathbf{E}}^{-1}(t) - g_{P_{\mathbf{E}\#}\mu}^{-1}(t)| dt \right)^{\frac{p}{q}} \leq C_2^{p/q} n^{p/2q}. \end{aligned}$$

When $p \geq q$, we have

$$\begin{aligned} & \mathbb{E} \left(\int_{\mathbf{E} \in \mathbb{S}_{d,q}} \frac{1}{n} \sum_{i=1}^n \int_{\frac{i-1}{n}}^{\frac{i}{n}} \|x_{\mathbf{E},(i)^*} - H_{P_{\mathbf{E}\#}\mu}(g_{P_{\mathbf{E}\#}\mu}^{-1}(t))\|_p^p dt d\sigma(\mathbf{E}) \right)^{\frac{1}{p}} \\ & \leq \left(\int_{\mathbf{E} \in \mathbb{S}_{d,q}} \mathbb{E} \frac{1}{n} \sum_{i=1}^n \int_{\frac{i-1}{n}}^{\frac{i}{n}} \|x_{\mathbf{E},(i)^*} - H_{P_{\mathbf{E}\#}\mu}(g_{P_{\mathbf{E}\#}\mu}^{-1}(t))\|_p^p dt d\sigma(\mathbf{E}) \right)^{\frac{1}{p}} \\ & \leq L n^{-\frac{1}{2p}}. \end{aligned}$$

When $p < q$, we have

$$\begin{aligned} & \mathbb{E} \left(\int_{\mathbf{E} \in \mathbb{S}_{d,q}} \frac{1}{n} \sum_{i=1}^n \int_{\frac{i-1}{n}}^{\frac{i}{n}} \|x_{\mathbf{E},(i)^*} - H_{P_{\mathbf{E}\#}\mu}(g_{P_{\mathbf{E}\#}\mu}^{-1}(t))\|_p^p dt d\sigma(\mathbf{E}) \right)^{\frac{1}{p}} \overline{\text{HCP}}_{p,k}(\mu, \mu_n) = \left(\int_0^1 \|H_\mu(g_\mu^{-1}(t)) - \widehat{H}'_{k,\mu}(\widehat{g}_{\mu_n,k}^{-1}(t))\|_p^p dt \right)^{\frac{1}{p}}. \end{aligned}$$

where L is a constant. Following the proof in Corollary 3.1, we have finished the proof. \square

Proof of Theorem 7. By the definition of PRHCP and Theorem 1, we have

$$\begin{aligned} \text{PRHCP}_{p,q}(\mu, \nu) &= \sup_{\mathbf{E} \in \mathbb{S}_{d,q}} \text{HCP}_p(P_{\mathbf{E}\#}\mu, P_{\mathbf{E}\#}\nu) \\ &\geq \sup_{\mathbf{E} \in \mathbb{S}_{d,q}} W_p(P_{\mathbf{E}\#}\mu, P_{\mathbf{E}\#}\nu) \\ &= \text{PRW}_{p,q}(\mu, \nu). \end{aligned}$$

By definition of PRHCP, it is clear that $\text{PRHCP}_{p,q}(\mu, \nu) \geq 0$ and $\text{PRHCP}_{p,q}(\mu, \nu) = \text{PRHCP}_{p,q}(\nu, \mu)$. Because $\text{PRW}_{p,q}$ is a well-defined metric, we can easily get $\text{PRHCP}_{p,q}(\mu, \nu) = 0$ if and only if $\mu = \nu$. Next, we prove the triangle inequality as follows,

$$\begin{aligned} & \text{PRHCP}_{p,q}(\mu, \nu) \\ &= \sup_{\mathbf{E} \in \mathbb{S}_{d,q}} \text{HCP}_p(P_{\mathbf{E}\#}\mu, P_{\mathbf{E}\#}\nu) \\ &\leq \sup_{\mathbf{E} \in \mathbb{S}_{d,q}} \text{HCP}_p(P_{\mathbf{E}\#}\mu, P_{\mathbf{E}\#}\nu) + \text{HCP}_p(P_{\mathbf{E}\#}\nu, P_{\mathbf{E}\#}\tau) \\ &\leq \sup_{\mathbf{E} \in \mathbb{S}_{d,q}} \text{HCP}_p(P_{\mathbf{E}\#}\mu, P_{\mathbf{E}\#}\nu) + \sup_{\mathbf{E} \in \mathbb{S}_{d,q}} \text{HCP}_p(P_{\mathbf{E}\#}\nu, P_{\mathbf{E}\#}\tau) \\ &\leq \text{PRHCP}_{p,q}(\mu, \tau) + \text{PRHCP}_{p,q}(\nu, \tau), \end{aligned}$$

where the first inequality uses triangle inequality from Theorem 1 and the second uses Minkowski inequality. \square

A.4 Statistical convergence analysis for k -order approximation Hilbert curve

The bound of convergence rate in our manuscript is regardless of k since we used the Hilbert curve rather than the k -order approximation Hilbert curve in the statistical convergence analysis. In practice, we always approximate the Hilbert curve using k -order Hilbert curve. Therefore, the bound should depend on the parameter k in practice. We now provide a more precise analysis in the following.

As shown in Section 3.4, following the definition in [57], the k -order Hilbert curves, i.e., $\widehat{H}_{k,\mu}$ partitions both $[0, 1]$

and $\widehat{\Omega}_\mu$ into 2^{kd} equal blocks, denoted by $\{c'_{j,\mu}\}_{j=1}^{2^{dk}}$ and $\{c_{j,\mu}\}_{j=1}^{2^{dk}}$, respectively, and construct a bijection between these blocks (this bijection is along Hilbert curve). Following the strategy used in the manuscript, we study a modified empirical k -order Hilbert curve projection distance instead. Specifically, following the definitions of g_μ , we first define the cumulative distribution function and its inverse for the empirical measure μ_n corresponding to the k -order Hilbert curve:

$$\begin{aligned} \widehat{g}_{\mu_n,k}(t) &= \inf_{s \in \mathcal{K}_k, s \geq t} \mu_n(\widehat{H}_{k,\mu}([0, s])), \\ \widehat{g}_{\mu_n,k}^{-1}(t) &= \inf_{s \in [0, 1], \widehat{g}_{\mu_n,k}(s) > t} s. \end{aligned}$$

where $\mathcal{K}_k = \{\frac{m_1}{2^{dk}} : m_1 \in \mathbb{N}, m_1 \leq 2^{dk}\}$. Accordingly, we define the modified empirical k -order Hilbert curve projection distance as:

$$\overline{\text{HCP}}_{p,k}(\mu, \mu_n) = \left(\int_0^1 \|H_\mu(g_\mu^{-1}(t)) - \widehat{H}'_{k,\mu}(\widehat{g}_{\mu_n,k}^{-1}(t))\|_p^p dt \right)^{\frac{1}{p}}.$$

where $\widehat{H}'_{k,\mu}(x)$ maps x in block $c'_{j,\mu}$ to the center of the block $c_{j,\mu}$. The following theorem provides an upper bound for the convergence rate.

Theorem 8. Let $\{x_i\}_{i=1}^n$ be an i.i.d. sample that is generated from the probability measure $\mu \in \mathcal{P}_\infty(\mathbb{R}^d)$. The empirical measure is defined by $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$. Then, we have

$$\begin{aligned} \mathbb{E} \overline{\text{HCP}}_{p,k}(\mu, \mu_n) &\lesssim \left(\max_{0 \leq i \leq 2^{dk}-1} \mu \left(H_\mu \left(\left[\frac{i}{2^{dk}}, \frac{i+1}{2^{dk}} \right] \right) \right) \right)^{\frac{1}{\max\{d,p\}}} \\ &\quad + O(n^{-\frac{1}{2\max\{p,d\}}}) + \frac{d^{1/p}}{2^k}. \end{aligned}$$

Proof. We know that

$$\begin{aligned} \overline{\text{HCP}}_{p,k}(\mu, \mu_n) &= \left(\int_0^1 \|H_\mu(g_\mu^{-1}(t)) - \widehat{H}'_{k,\mu}(\widehat{g}_{\mu_n,k}^{-1}(t))\|_p^p dt \right)^{\frac{1}{p}} \\ &\leq \left(\int_0^1 \|H_\mu(g_\mu^{-1}(t)) - H_\mu(\widehat{g}_{\mu_n,k}^{-1}(t))\|_p^p dt \right)^{\frac{1}{p}} \\ &\quad + \left(\int_0^1 \|H_\mu(\widehat{g}_{\mu_n,k}^{-1}(t)) - \widehat{H}'_{k,\mu}(\widehat{g}_{\mu_n,k}^{-1}(t))\|_p^p dt \right)^{\frac{1}{p}} \\ &\lesssim \left(\int_0^1 |g_\mu^{-1}(t) - \widehat{g}_{\mu_n,k}^{-1}(t)|^{\frac{p}{d}} dt \right)^{\frac{1}{p}} \\ &\quad + \left(\int_0^1 \|H_\mu(\widehat{g}_{\mu_n,k}^{-1}(t)) - \widehat{H}'_{k,\mu}(\widehat{g}_{\mu_n,k}^{-1}(t))\|_p^p dt \right)^{\frac{1}{p}} \end{aligned}$$

where the first inequality uses Minkowski inequality and triangle inequality (similar to the proof of Theorem 1) and the second inequality uses locality-preserving property. Since $H_\mu(t)$ and $\widehat{H}'_{k,\mu}(t)$ lie in the same block, it is easy to show that $\|H_\mu(t) - \widehat{H}'_{k,\mu}(t)\|_p^p \leq \frac{d}{2^{pk}}$ for any given $t \in [0, 1]$.

We have when $p \geq d$

$$\left(\int_0^1 |g_\mu^{-1}(t) - \widehat{g}_{\mu_n,k}^{-1}(t)|^{\frac{p}{d}} dt \right)^{\frac{1}{p}} \leq \left(\int_0^1 |g_\mu^{-1}(t) - \widehat{g}_{\mu_n,k}^{-1}(t)| dt \right)^{\frac{1}{d}}.$$

When $p < d$, by Jensen's inequality, we have

$$\left(\int_0^1 |g_\mu^{-1}(t) - \widehat{g}_{\mu_n,k}^{-1}(t)|^{\frac{p}{d}} dt \right)^{\frac{1}{p}} \leq \left(\int_0^1 |g_\mu^{-1}(t) - \widehat{g}_{\mu_n,k}^{-1}(t)| dt \right)^{\frac{1}{d}}.$$

And we have $|g_\mu^{-1}(t) - \hat{g}_{\mu_n,k}^{-1}(t)| \leq |g_\mu^{-1}(t) - g_{\mu,k}^{-1}(t)| + |g_{\mu,k}^{-1}(t) - \hat{g}_{\mu_n,k}^{-1}(t)|$, where

$$\begin{aligned} g_{\mu,k}(t) &= \inf_{s \in \mathcal{K}_k, s \geq t} \mu(\hat{H}_{k,\mu}([0, s])), \\ g_{\mu,k}^{-1}(t) &= \inf_{s \in [0, 1], g_{\mu,k}(s) > t} s. \end{aligned}$$

Thus, we know that

$$\begin{aligned} &\mathbb{E}\overline{\text{HCP}}_{p,k}(\mu, \mu_n) \\ &\lesssim \left(\int_0^1 |g_\mu^{-1}(t) - g_{\mu,k}^{-1}(t)| dt \right)^{\frac{1}{\max\{d,p\}}} \\ &+ \left(\mathbb{E} \int_0^1 |g_{\mu,k}^{-1}(t) - \hat{g}_{\mu_n,k}^{-1}(t)| dt \right)^{\frac{1}{\max\{d,p\}}} + \frac{d^{1/p}}{2^k} \\ &\lesssim \left(\int_0^1 |g_\mu^{-1}(t) - g_{\mu,k}^{-1}(t)| dt \right)^{\frac{1}{\max\{d,p\}}} + O(n^{-\frac{1}{2\max\{p,d\}}}) + \frac{d^{1/p}}{2^k} \\ &= \left(\int_0^1 |g_\mu(t) - g_{\mu,k}(t)| dt \right)^{\frac{1}{\max\{d,p\}}} + O(n^{-\frac{1}{2\max\{p,d\}}}) + \frac{d^{1/p}}{2^k} \\ &\leq \left(\max_i \mu(H_\mu([\frac{i}{2^k}, \frac{i+1}{2^k}])) \right)^{\frac{1}{\max\{d,p\}}} + O(n^{-\frac{1}{2\max\{p,d\}}}) + \frac{d^{1/p}}{2^k} \end{aligned}$$

where the second inequality is from [63] (similar to the proof of Theorem 3), the third equality is from [64] and the last inequality is from the fact that $|g_\mu(t) - g_{\mu,k}(t)| \leq \max_i \mu(H_\mu([\frac{i}{2^k}, \frac{i+1}{2^k}]))$. \square

Remark 1. If μ is absolutely continuous with bounded density function f , then we have

$$\mathbb{E}\overline{\text{HCP}}_{p,k}(\mu, \mu_n) \lesssim \left(\frac{C}{2^k} \right)^{\frac{1}{\max\{d,p\}}} + O(n^{-\frac{1}{2\max\{p,d\}}}) + \frac{d^{1/p}}{2^k}.$$

for some constant $C > 0$.

The above theoretical analysis shows that we could take $k = O(\log(n))$ in practice.

REFERENCES

- [1] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR, 2015.
- [2] Gao Huang, Chuan Guo, Matt J Kusner, Yu Sun, Fei Sha, and Kilian Q Weinberger. Supervised word mover’s distance. *Advances in neural information processing systems*, 29, 2016.
- [3] Alain Rakotomamonjy, Abraham Traoré, Maxime Berar, Rémi Flamary, and Nicolas Courty. Distance measure machines. *arXiv preprint arXiv:1803.00250*, 2018.
- [4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [5] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014.
- [6] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.
- [7] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.
- [8] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.
- [9] I Tolstikhin, O Bousquet, S Gelly, and B Schölkopf. Wasserstein auto-encoders. In *6th International Conference on Learning Representations*, 2018.
- [10] Yann Brenier. A homogenized model for vortex sheets. *Archive for Rational Mechanics and Analysis*, 138(4):319–353, 1997.
- [11] J-D Benamou, Yann Brenier, and Kevin Guittet. The Monge–Kantorovitch mass transfer and its computational fluid mechanics formulation. *International Journal for Numerical methods in fluids*, 40(1-2):21–30, 2002.
- [12] Yossi Rubner, Leonidas J Guibas, and Carlo Tomasi. The earth mover’s distance, multi-dimensional scaling, and color-based image retrieval. In *Proceedings of the ARPA image understanding workshop*, volume 661, page 668, 1997.
- [13] Ofir Pele and Michael Werman. Fast and robust earth mover’s distances. In *2009 IEEE 12th International Conference on Computer Vision*, pages 460–467. IEEE, 2009.
- [14] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26:2292–2300, 2013.
- [15] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.
- [16] Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and K Gustavo. Generalized sliced Wasserstein distances. *Advances in Neural Information Processing Systems*, 2019.
- [17] Tam Le, Makoto Yamada, Kenji Fukumizu, and Marco Cuturi. Tree-sliced variants of Wasserstein distances. In *Advances in Neural Information Processing Systems*, 2019.
- [18] Michael Bader. *Space-filling curves: an introduction with applications in scientific computing*. Springer, 2012.
- [19] David J Abel and David M Mark. A comparative analysis of some two-dimensional orderings. *International Journal of Geographical Information Systems*, 4(1):21–31, 1990.
- [20] Bongki Moon, Hosagrahar V Jagadish, Christos Faloutsos, and Joel H. Saltz. Analysis of the clustering properties of the Hilbert space-filling curve. *IEEE Transactions on knowledge and data engineering*, 13(1):124–141, 2001.
- [21] Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G Schwing. Max-sliced Wasserstein distance and its use for GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10648–10656, 2019.
- [22] Soheil Kolouri, Phillip E Pope, Charles E Martin, and Gustavo K Rohde. Sliced Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.
- [23] Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. *Advances in*

- neural information processing systems*, 30, 2017.
- [24] Jason Altschuler, Francis Bach, Alessandro Rudi, and Jonathan Niles-Weed. Massively scalable Sinkhorn distances via the Nyström method. *Advances in neural information processing systems*, 32, 2019.
- [25] Pavel Dvurechensky, Alexander Gasnikov, Sergey Omelchenko, and Alexander Tiurin. Adaptive similar triangles method: A stable alternative to Sinkhorn’s algorithm for regularized optimal transport. *arXiv preprint arXiv:1706.07622*, 2017.
- [26] Alexis Thibault, Lénaïc Chizat, Charles Dossal, and Nicolas Papadakis. Overrelaxed Sinkhorn–Knopp algorithm for regularized optimal transport. *Algorithms*, 14(5):143, 2021.
- [27] Mengyu Li, Jun Yu, Tao Li, and Cheng Meng. Importance sparsification for Sinkhorn algorithm. *Journal of Machine Learning Research*, 24:1–44, 2023.
- [28] Qichen Liao, Jing Chen, Zihao Wang, Bo Bai, Shi Jin, and Hao Wu. Fast Sinkhorn I: An $O(N)$ algorithm for the Wasserstein-1 metric. *Communications in Mathematical Sciences*, 20(7):2053–2067, 2022.
- [29] Wenshuo Guo, Nhat Ho, and Michael Jordan. Fast algorithms for computational optimal transport and Wasserstein barycenter. In *International Conference on Artificial Intelligence and Statistics*, pages 2088–2097. PMLR, 2020.
- [30] Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn’s algorithm. In *International conference on machine learning*, pages 1367–1376. PMLR, 2018.
- [31] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. *Advances in neural information processing systems*, 29, 2016.
- [32] Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha. A fast proximal point method for computing exact Wasserstein distance. In *Uncertainty in artificial intelligence*, pages 433–453. PMLR, 2020.
- [33] Huahua Wang and Arindam Banerjee. Bregman alternating direction method of multipliers. *Advances in Neural Information Processing Systems*, 27, 2014.
- [34] Jianbo Ye, Panruo Wu, James Z Wang, and Jia Li. Fast discrete distribution clustering using Wasserstein barycenter with sparse support. *IEEE Transactions on Signal Processing*, 65(9):2317–2332, 2017.
- [35] Hongteng Xu, Jiachang Liu, Dixin Luo, and Lawrence Carin. Representing graphs via Gromov-Wasserstein factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [36] Khai Nguyen, Nhat Ho, Tung Pham, and Hung Bui. Distributional sliced-Wasserstein and applications to generative modeling. In *International Conference on Learning Representations*, 2020.
- [37] Mark Rowland, Jiri Hron, Yunhao Tang, Krzysztof Choromanski, Tamas Sarlos, and Adrian Weller. Orthogonal estimation of Wasserstein distances. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 186–195. PMLR, 2019.
- [38] François-Pierre Paty and Marco Cuturi. Subspace robust Wasserstein distances. In *International conference on machine learning*, pages 5072–5081. PMLR, 2019.
- [39] Tianyi Lin, Chenyou Fan, Nhat Ho, Marco Cuturi, and Michael Jordan. Projection robust Wasserstein distance and Riemannian optimization. *Advances in neural information processing systems*, 33:9383–9397, 2020.
- [40] Tianyi Lin, Zeyu Zheng, Elynn Chen, Marco Cuturi, and Michael I Jordan. On projection robust optimal transport: Sample complexity and model misspecification. In *International Conference on Artificial Intelligence and Statistics*, pages 262–270. PMLR, 2021.
- [41] Khai Nguyen and Nhat Ho. Revisiting sliced Wasserstein on images: From vectorization to convolution. *arXiv preprint arXiv:2204.01188*, 2022.
- [42] Khai Nguyen and Nhat Ho. Amortized projection optimization for sliced Wasserstein generative models. *arXiv preprint arXiv:2203.13417*, 2022.
- [43] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR, 2018.
- [44] Jiqing Wu, Zhiwu Huang, Dinesh Acharya, Wen Li, Janine Thoma, Danda Pani Paudel, and Luc Van Gool. Sliced Wasserstein generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3713–3722, 2019.
- [45] Hongteng Xu, Dixin Luo, Ricardo Henao, Svati Shah, and Lawrence Carin. Learning autoencoders with relational regularization. In *International Conference on Machine Learning*, pages 10576–10586. PMLR, 2020.
- [46] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a Wasserstein loss. *Advances in neural information processing systems*, 28, 2015.
- [47] Matteo Togninalli, Elisabetta Ghisu, Felipe Llinares-López, Bastian Rieck, and Karsten Borgwardt. Wasserstein Weisfeiler-Lehman graph kernels. *Advances in Neural Information Processing Systems*, 32, 2019.
- [48] Mengyu Li, Jun Yu, Hongteng Xu, and Cheng Meng. Efficient approximation of Gromov-Wasserstein distance using importance sparsification. *Journal of Computational and Graphical Statistics*, pages 1–12, 2023.
- [49] Cheng Meng, Jun Yu, Jingyi Zhang, Ping Ma, and Wenxuan Zhong. Sufficient dimension reduction for classification using principal optimal transport direction. *Advances in Neural Information Processing Systems*, 33:4015–4028, 2020.
- [50] Rémi Flamary, Marco Cuturi, Nicolas Courty, and Alain Rakotomamonjy. Wasserstein discriminant analysis. *Machine Learning*, 107(12):1923–1945, 2018.
- [51] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in Neural Information Processing Systems*, 30, 2017.
- [52] Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin Duke. Gromov-Wasserstein learning for graph matching and node embedding. In *International conference on machine learning*, pages 6932–6941. PMLR, 2019.
- [53] Hongteng Xu, Dixin Luo, and Lawrence Carin. Scalable Gromov-Wasserstein learning for graph partitioning and matching. *Advances in neural information processing systems*, 33:4015–4028, 2020.

- systems*, 32, 2019.
- [54] Julien Rabin, Sira Ferradans, and Nicolas Papadakis. Adaptive color transfer with relaxed optimal transport. In *2014 IEEE international conference on image processing (ICIP)*, pages 4852–4856. IEEE, 2014.
- [55] Cheng Meng, Yuan Ke, Jingyi Zhang, Mengrui Zhang, Wenxuan Zhong, and Ping Ma. Large-scale optimal transport map estimation using projection pursuit. *Advances in Neural Information Processing Systems*, 32, 2019.
- [56] Mikhail Yurochkin, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh, and Justin M Solomon. Hierarchical optimal transport for document representation. *Advances in Neural Information Processing Systems*, 32, 2019.
- [57] Zhijian He and Art B Owen. Extensible grids: uniform sampling on a space filling curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):917–931, 2016.
- [58] Gerhard Zumbusch. *Parallel multilevel methods: adaptive mesh refinement and loadbalancing*. Springer Science & Business Media, 2012.
- [59] Kilian Fatras, Younes Zine, Rémi Flamary, Remi Gribonval, and Nicolas Courty. Learning with minibatch Wasserstein: asymptotic and gradient properties. In *International Conference on Artificial Intelligence and Statistics*, pages 2131–2141. PMLR, 2020.
- [60] Benedetto Piccoli and Francesco Rossi. Generalized Wasserstein distance and its application to transport equations with source. *Archive for Rational Mechanics and Analysis*, 211:335–358, 2014.
- [61] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.
- [62] Jiaqi Xi and Jonathan Niles-Weed. Distributional convergence of the sliced Wasserstein process. *Advances in Neural Information Processing Systems*, 35:13961–13973, 2022.
- [63] Victor M Panaretos and Yoav Zemel. Statistical aspects of Wasserstein distances. *Annual review of statistics and its application*, 6:405–431, 2019.
- [64] Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.
- [65] Arthur R Butz. Convergence with Hilbert’s space-filling curve. *Journal of Computer and System Sciences*, 3(2):128–146, 1969.
- [66] Arthur R Butz. Alternative algorithm for Hilbert’s space-filling curve. *IEEE Transactions on Computers*, 100(4):424–426, 1971.
- [67] John Skilling. Programming the Hilbert curve. In *AIP Conference Proceedings*, volume 707, pages 381–387. American Institute of Physics, 2004.
- [68] A Tanaka. *Study on a fast ordering of high dimensional data to spatial index*. PhD thesis, Master thesis, Kyushu Institute of Technology, 2001.
- [69] Yasunobu Imamura, Takeshi Shinohara, Kouichi Hirata, and Tetsuji Kuboyama. Fast Hilbert sort algorithm without using Hilbert indices. In *International Conference on Similarity Search and Applications*, pages 259–267. Springer, 2016.
- [70] Andreas Fabri and Sylvain Pion. CGAL: The computational geometry algorithms library. In *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 538–539, 2009.
- [71] Chris H Hamilton and Andrew Rau-Chaplin. Compact Hilbert indices: Space-filling curves for domains with unequal side lengths. *Information Processing Letters*, 105(5):155–163, 2008.
- [72] Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(2):235–269, 2019.
- [73] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.
- [74] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.
- [75] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.
- [76] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [77] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *International Conference on Learning Representations*, 2016.
- [78] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.
- [79] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [80] Yogesh Balaji, Rama Chellappa, and Soheil Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. *Advances in Neural Information Processing Systems*, 33:12934–12944, 2020.
- [81] Debarghya Mukherjee, Aritra Guha, Justin M Solomon, Yuekai Sun, and Mikhail Yurochkin. Outlier-robust optimal transport. In *International Conference on Machine Learning*, pages 7850–7860. PMLR, 2021.
- [82] Khang Le, Huy Nguyen, Quang M Nguyen, Tung Pham, Hung Bui, and Nhat Ho. On robust optimal transport: Computational complexity and barycenter computation. *Advances in Neural Information Processing Systems*, 34:21947–21959, 2021.
- [83] Khiem Pham, Khang Le, Nhat Ho, Tung Pham, and Hung Bui. On unbalanced optimal transport: An analy-

- sis of Sinkhorn algorithm. In *International Conference on Machine Learning*, pages 7673–7682. PMLR, 2020.
- [84] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.
- [85] Thibault Séjourné, Clément Bonet, Kilian Fatras, Kimia Nadjahi, and Nicolas Courty. Unbalanced optimal transport meets sliced-Wasserstein. *arXiv preprint arXiv:2306.07176*, 2023.
- [86] Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.
- [87] Brendan Pass. Multi-marginal optimal transport: theory and applications. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 49(6):1771–1790, 2015.
- [88] Isabel Haasler, Rahul Singh, Qinsheng Zhang, Johan Karlsson, and Yongxin Chen. Multi-marginal optimal transport and probabilistic graphical models. *IEEE Transactions on Information Theory*, 67(7):4647–4668, 2021.
- [89] Martial Aguech and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.
- [90] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov–Wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672. PMLR, 2016.
- [91] Justin Luitjens, Martin Berzins, and Tom Henderson. Parallel space-filling curve generation through sorting. *Concurrency and Computation: Practice and Experience*, 19(10):1387–1402, 2007.
- [92] Ibrahim Kamel and Christos Faloutsos. Hilbert R-tree: An improved R-tree using fractals. Technical report, 1993.
- [93] Nicolas Bonnotte. *Unidimensional and evolution methods for optimal transportation*. PhD thesis, Paris 11, 2013.
- [94] Sergey Bobkov and Michel Ledoux. *One-dimensional empirical measures, order statistics, and Kantorovich transport distances*, volume 261. American Mathematical Society, 2019.



Tao Li is a Ph.D. student in the Institute of Statistics and Big Data, Renmin University of China. He received his BS degree in mathematics from Nanjing University in 2019. His research interests include optimal transport problems, generative model, sufficient dimension reduction and variable selection.



Cheng Meng is an assistant professor (Tenure-Track) in the Institute of Statistics and Big Data, Renmin University of China. He received his Ph.D. from the Department of Statistics at University of Georgia in 2020. His research interests include numerical linear algebra, optimal transport problems, sufficient dimension reduction, nonparametric statistics, and machine learning.



Hongteng Xu is an associate professor (Tenure-Track) in the Gaoling School of Artificial Intelligence, Renmin University of China. From 2018 to 2020, he was a senior research scientist in Infinia ML Inc. In the same time period, he is a visiting faculty member in the Department of Electrical and Computer Engineering, Duke University. He received his Ph.D. from the School of Electrical and Computer Engineering at Georgia Institute of Technology (Georgia Tech) in 2017. His research interests include machine learning and its applications, especially optimal transport theory, sequential data modeling and analysis, deep learning techniques, and their applications in computer vision and data mining.



Jun Yu is an associate professor in the school of mathematics and statistics, Beijing Key Laboratory on MCAACI, and key laboratory of mathematical theory and computation in information security, Beijing institute of technology. He received the Ph.D degree in statistics from Peking University, China in 2019. His research interests include the design of experiments, statistical sketching and sampling methods for large-scale data, and applied statistics in solving scientific and engineering problems.